NWO
Netherlands Organisation for Scientific Research

Council for Physical Sciences
Council for Humanities

CATCH

A computer science research programme for Continuous Access To Cultural Heritage

November 2004

This text has been written by the CATCH Programme Preparation Committee:

Prof.dr. H.J. van den Herik	chair	computer science	UM
Drs. P.M. Doorenbosch	vice chair	cultural heritage	KB
Prof.dr. F.A.H. van Harmelen	member	computer science	VU
Drs. J.Th. Taekema	member	cultural heritage	DEN
Dr. A.P.J. van den Bosch	member	computer science	UvT
Ir. D.G. Houtgraaf	member	cultural heritage	Naturalis
Dr. P.K. Doorn	member	information science	KNAW
Drs. A.M. Bos	member	NWO Humanities	
Dr. A.P. Meijler	member	NWO Physical Sciences	
Drs. A. Dijkstra	coordinator	NWO Humanities	
Dr. M. Kas	coordinator	NWO Physical Sciences	

TABLE OF CONTENTS

SUMMARY

The collective memory of the Netherlands is stored in our cultural heritage. The total size of the Dutch cultural heritage is certain to be huge. In the Netherlands there are at least 80 large collections that together contain more than several millions of objects. The economic value of this heritage (estimated at 22 billion euros, art collections only) underscores the enormous value of our cultural heritage. Cultural heritage belongs to the entire population of our country and plays a role in many aspects of society: tourism, education, research, cultural interest etc.

For historical reasons, the collections of physical objects have landed in a large number of cultural heritage institutions. This poses limitations for both visitors and researchers. Digitisation holds the promise for continuous access to all cultural heritage collections, unrestricted by time and space. All the digitised collections of the cultural heritage institutes form one large Ambient Heritage Collection. This opens unimagined possibilities for research, education, cultural leisure, and tourism.

Despite large investments, the cultural heritage institutions encounter a number of persistent obstacles that are hindering progress. There is a strong sense of urgency felt by the organisations in the cultural heritage domain to come up with new solutions to get access to the data of the digitised collections. The volume of the Dutch cultural heritage is immense and increasing everyday. A new approach has to be developed. The CATCH programme aims to do research in order to find these new solutions. The two central research questions in CATCH are:

- To what extent is it possible to develop innovative tools (1) to connect knowledge and cultural objects, (2) to integrate scattered digitised cultural objects and (3) to increase the accessibility of and the interaction with our cultural heritage supporting and improving the work of the professionals?
- Can we develop scientifically relevant methods to acquire new fundamental and applied knowledge about these processes and their IT-based solutions?

The challenges implied by the research questions are common to all cultural heritage institutions in the world. The CATCH programme joins the ongoing international efforts. On the one hand CATCH aims to develop tools to improve the specific situation for Dutch cultural heritage (research question 1). On the other hand CATCH wants to contribute new methods and techniques to the international research effort (research question 2).

The CATCH research goals have been established in a process that can be characterised as *demand pull* rather than *technology push*. In a demand-pull programme the interests of the (potential) users of the research results are of outstanding importance. Hence the programme strategy has a twofold focus: research and implementation. As a direct consequence, the CATCH programme will have two types of results:

- new knowledge
- software (tools).

The challenges for the CATCH programme: (1) multidisciplinary cooperation between cultural heritage and IT research, (2) excellent research contributions, and (3) intelligent and personalised tools. The CATCH research strategy concentrates on three research themes.

THEME 1: Semantic interoperability through metadata
THEME 2: Knowledge enrichment through automated analyses
THEME 3: Personalisation through presentation

The CATCH research focuses on the development of tools and methods to speed up the back office processes, i.e. tools and methods that will enable the collection managers of the cultural heritage institutes to do more in less time and with higher quality. All developed tools and algorithms will be implemented in two 'integrators', existing large IT-projects of national importance in the cultural heritage field.

CATCH is a coordinated effort with respect to three strategies: research, implementation and support.

The research and implementation will be done by research teams consisting of CATCH-funded temporary researchers (PhD students, postdocs), temporary scientific programmers and senior research staff (all employed by universities), and programmers and senior staff employed by cultural heritage institutions (researchers and/or collection managers or others with relevant expertise). With an estimated total budget of M€ 12,5 in subsidies (to be realised in to phases), CATCH will be able to fund about 17 of these research teams. The programme will start with six research teams, each executing one of the six core projects which lay the foundation for the programme. The 11 remaining teams will be selected in competition on the basis of research plans. All Dutch universities can enter the competition, which will be organised by NWO. The participating cultural heritage institutions will contribute M€ 2,8 in kind to the programme.

The support programme provides for the transfer of knowledge and tools (a) within the programme and (b) to all other parties interested in the CATCH results. Furthermore, the support programme aims at building and establishing a structure which guarantees continuity for the results (in particular the tools, the software, and the knowledge) of the CATCH programme.

The programme will be run by a Programme Committee with representatives of the three CATCH themes and additional experts. Daily affairs will be taken care of by an Executive Committee and the Programme Management Bureau. A Steering Committee representing all parties contributing financially to the programme is responsible for the supervision of the programme and all major (financial) decisions. Programme Committee and Steering Committee are assisted by an International Scientific Advisory Board.

The CATCH programme starts in November 2004 and will run for six years.

1. GENERAL PROBLEM STATEMENT

The collective memory of the Netherlands is stored in our cultural heritage. Enormous amounts of archives, books and magazines, paintings and other objects of art, audiovisual sources, objects of folklore, archaeological remains, and logs describing these objects are kept in numerous places, often in buildings that form part of our cultural heritage themselves. The total size of the Dutch cultural heritage is difficult to estimate but is certain to be huge. In the Netherlands there are at least 80 large collections that together contain more than several millions of objects.[1] The economic value of this heritage is even more difficult to estimate since the true value is symbolic rather than economic. Nevertheless, the estimated monetary value (22 billion euros) (1998[2], art collections only) underscores the enormous value of our cultural heritage. This is accentuated by the fact that the government is spending around 200 to 250 million euro on an annual basis on the management of the cultural-heritage sector. Revenues and secondary economic effects are probably much larger.

All these witnesses of our past and present are indispensable components of our national identity. Cultural heritage belongs to the entire population of our country and plays a role in many aspects of society: tourism, education, research, cultural interest etc. For historical reasons, the collections of physical objects have landed in a large number of cultural heritage institutions. This poses limitations for both visitors and researchers. Related objects are often stored at different locations. For centuries these limitations were overcome through physical movement. Visitors and researchers travelled to the objects they desired to see, or related objects belonging to different collections were moved to one place to form an exhibition. Yet because of the limitations of time and space the accessibility remained inherently restricted.

Digitisation holds the promise for continuous access to all cultural heritage collections, unrestricted by time and space. Physical constraints no longer apply. All the digitised collections of the cultural heritage institutes form one large Ambient Heritage Collection. This opens unimagined possibilities for research, education, cultural leisure, and tourism. The cultural heritage institutions and the government are very much aware of the potential possibilities the new information technology offers them to perform their public tasks: to preserve, present and propagate their collections to audiences ranging from specialised researchers to the general public. They invest heavily in the digitisation of their collections and the accessibility of the collections through the internet. There are a number of excellent examples where large digital collections have been made available to large audiences.

Despite these investments and other major efforts, the cultural heritage institutions encounter a number of persistent obstacles that are hindering progress. Below they are summarised in five points.

1. *The digitisation process is slow*, often cumbersome, and therefore very expensive. Most heritage objects are precious and have to be handled with care. Refined technical solutions are needed to support and automate the digitisation process with the subtlety required by such precious goods.

[1] *Quick scan Digitalisering Cultureel Erfgoed in Nederlandse Collecties. Reekx Advies, April 2002.*

[2] *Source: CBS.*

2. *Independent collections, unconnected databases*. In the same way as physical objects are kept in numerous independent collections, their digital counterparts are stored in a huge archipelago of (more or less) unconnected databases. Connecting these databases and making them interoperable is a complicated problem, which needs to be solved if the promises to lift the limitations of time and space are ever to be fulfilled.
3. *Access problems*. Even if the databases are technically connected and can be approached as though they were one large system, there remains the problem to search and sift through millions and millions of objects, ranging from written text to spoken text, from still images to moving images, from 2D objects to 3D objects, and to find the objects one was looking for. Progress is hampered by the great variety of schemes and systems describing the semantics of the objects.
4. *The problem of knowledge enrichment*. Finding the objects, however, is not enough if we want to exploit the potential of the new digital world to the largest extent possible. Data from various sources (e.g., text and images) can be connected in sensible ways to give us deeper insight into the nature of objects (e.g., paintings) or processes (e.g., historical events). The challenge is to find automated ways to make new knowledge out of existing data and knowledge.
5. *The problem of personalisation*. The results of the searches have to be presented in ways that correspond to the needs of the person who was looking for the information. It is almost trivial to remark that the presentation of the results of a search to a specialised researcher can, and probably have to, be of another nature than the presentation of the same results to an eight-year-old child. However, it is far from trivial to devise the techniques to realise this.

There is a strong sense of urgency felt by the organisations in the cultural heritage domain to come up with new solutions to get access to the data. The volume of the Dutch cultural heritage is immense and increasing everyday. The funds and time required to be able to digitise and present all our cultural material in a traditional way are lacking by any means. Therefore, a new approach has to be developed since there is an increasing demand, stimulated by the use of internet.

This brings us to two central research questions.

- To what extent is it possible to develop innovative tools (1) to connect knowledge and cultural objects, (2) to virtually integrate scattered digitised cultural objects and (3) to increase the accessibility of and the interaction with our cultural heritage supporting and improving the work of the professionals?
- Can we develop scientifically relevant methods to acquire new fundamental and applied knowledge about these processes and their IT-based solutions?

The challenges implied by the research questions are common to all cultural heritage institutions in the world. Therefore, all over the world serious research efforts are realised to contribute to new ways of dealing with our cultural heritage. The CATCH programme joins these efforts. On the one hand CATCH aims to develop tools to improve the specific situation for Dutch cultural heritage (research question 1). On the other hand CATCH wants to contribute new methods and techniques to the international research effort (research question 2).

2. PROGRAMME STRATEGY

Essential in the CATCH research programme is the direct involvement of the cultural heritage sector in defining the aims and content of the research, right from the start. The CATCH research goals have been established in a process that can be characterised as *demand pull* rather than *technology push*. In a demand-pull programme the interests of the (potential) users of the research results are leading. The programme strategy – guided by the CATCH principle of interaction and cooperation - has a twofold focus: research and implementation. As a direct consequence, the CATCH programme will have two types of results:

- new knowledge
- software(tools)

A main characteristic of the CATCH programme is that the production of these two types of results is interwoven. Obviously, from a scientific point of view, IT-research has as its principal aim the development of new methods, techniques, insights, and knowledge. The results achieved can be equally beneficial for the cultural heritage sector as for the IT-research itself and a variety of commercial applications. Of course, all results will be disseminated, too, by papers, articles, dissertations etc. The universities and research institutions are responsible for the dissemination and preservation of this knowledge. Cultural heritage institutions and the participating companies should be able to have free access to the knowledge developed. Section 2.1 describes the programme's research strategy to produce new knowledge. Section 2.2 describes the programme's implementation strategy.

2.1 Research Strategy

Although the CATCH programme is ambitious, it has by no means the aspiration to deal with all obstacles mentioned in the previous chapter. Through a concerted and focused research effort, embedded within and guided by the leading Dutch cultural heritage institutions, CATCH aims at a measurable and permanent impact on an improved accessibility of digital cultural heritage.

Four characteristics of cultural heritage are particularly relevant to the CATCH programme.

1. The **volume** of the cultural heritage is **huge**.
2. The cultural-heritage objects are **distributed** over many distinct collections. They are exhibited or stored in 900 museums, 400 archives, and 1100 libraries in the Netherlands.
3. The collection of cultural-heritage objects is **heterogeneous**, ranging from buildings to books and pictures.
4. Cultural heritage is generated in a largely **unpredictable autonomous process**. Material and immaterial products of human activity and creativity enter the domain of cultural heritage in a continuous and perennial stream.

These characteristics combined with the obstacles mentioned earlier define the challenges for the CATCH programme: (1) multidisciplinary cooperation between cultural heritage and

IT research, (2) excellent research contributions, and (3) intelligent and personalised tools. The CATCH research strategy concentrates on three research themes.

THEME 1: Semantic interoperability through metadata
THEME 2: Knowledge enrichment through automated analyses
THEME 3: Personalisation through presentation

2.1.1 Theme 1: Semantic interoperability through metadata

Situation in cultural heritage

From the start, the cultural heritage institutes have used registration systems to add metadata to their collections. However, each of the highly autonomous institutes has done so in its own way. Only recently the institutes have become more aware of the need for standards in the structure of the descriptions, the conventions within the descriptions, and the terminological sources. Nowadays, the sheer amount of heritage sources, their great diversity, the amount of different registration systems used, and the ever evolving wishes of the users make it impossible to provide the "Dutch Heritage Collection" with unambiguous metadata through intellectual human labour. The challenge is to achieve the desired situation by combining intelligent IT applications and human expertise.

Hence, cultural heritage may turn to information technology with a clear technology demand for tools and methods (1) to combine and enrich the already registered data and knowledge, (2) to document sources automatically or semi-automatically, and (3) to supply them with the necessary metadata. The (semi-)automatic generation of metadata is an essential prerequisite for the semantic interoperability of the collections. Metadata not only makes sure that a person can find a specific collection or object, it also enables bulk retrieval of digital objects that are related to each other (e.g., created by the same artist, about the same topic, from the same period, from the same geographic location, etc.). Here we reiterate that the creation of such metadata usually requires a considerable intellectual input of curators and others involved in digital heritage collections. Information technology may offer opportunities for semantic interoperability between digital collections and their metadata on a large scale, which could not be achieved by human input alone. Finally, it is remarked that the creation of a Semantic Web can only be achieved by extensive IT research on semantic interoperability.

Research topics

The leading question is: How can we achieve the creation of semantic metadata by applying automatic creation of metadata? An obvious research agenda reads: (1) by deriving metadata from other collections, and (2) by using ontologies for adding additional elements in metadata corpora to guarantee 'semantic cohesion' between collections and items. Although the main goal is to provide methods and tools that can be used in the "back office" to create semantically rich metadata, there are two more questions, viz. on the speed of the project execution, and on the open structure of the solutions. The tools should minimize the amount of user effort required for creating and maintaining semantic annotations and should help to increase the overall quality level of annotations.

Research will focus on methods and tools for harmonizing ontologies through semantic links between metadata corpora. This research challenge is similar to what is called the "ontology mapping" problem. Research issues with respect to ontology mapping include the following five different topics.

- Inventory of (the composition of) ontologies and vocabularies that are of potential use for cultural heritage applications.
- Types of mapping relations: e.g., equality, equivalence, subclass, instance.
- Methods for representation of mapping relations: e.g., how to add mappings without affecting the original metadata vocabularies.
- Semi-automatic learning of mapping relations; techniques such as emergent semantics (learning semantic relations from user behaviour) may be relevant here.
- Methods for combining metadata with full text documents within a single query.

Background

To understand the research question and the research topics more in depth, we provide some background. The first two bullets underline the importance of metadata once more. The bullets three to five emphasize the various difficulties with semantics.

- Metadata can refer to various kinds of data types. It turns out that the limited and well-defined semantic scope of keyword type of metadata (like IMDI) can be seen as the backbone for collection maintenance and discovery.
- Keyword type of metadata is also one of the keys for interoperability due to the broad usage (community agreed on elements and use the same concepts) and well-defined limited semantics.
- Achieving semantic interoperability is a hard process where the goals have to be clear. The experience shows that most relationships between the elements of two disciplines can only be expressed with the help of a fuzzy type such as "mapsTo". Frameworks such as RDF(S) and OWL do not include such a relation type for good reasons. Actually, the "mapsTo" relation is exploited as a one-directional equality with some further necessary restrictions.
- The limited semantics of the keyword type of metadata and the fact that metadata creation is an expensive endeavour leading to missing values makes it necessary to use all types of contextual information (within metadata hierarchies/environments and outside) to enrich the metadata and to add it to the discovery domain. Both topics are completely new and not sorted out very well. Research has to be done to understand what is possible and how the quality of the metadata will be influenced. Also it has to be understood how metadata and context information can be combined to increase the chance of discovery.
- Semantic annotation has to rely on well-defined domain knowledge to form a coherent discovery space. Therefore, the concepts to be used should be taken from open data category registries (DCR). If a new concept is introduced due to the fact that the existing ones are semantically not sufficient, then the person intending to use it has the duty to enter it into the data category repository, i.e., defining it properly and also where possible define relationships with other existing concepts. The DCRs are essential to avoid a proliferation of concepts which would reduce its relevance for the discovery space and for achieving interoperability.

2.1.2 Theme 2: Knowledge enrichment through automated analyses

Situation in cultural heritage
Collection management and research in the cultural heritage field centres around content, i.e., the meaning of texts, objects, images and their mutual relations. For unanalysed objects, this information is hidden and implicit. The goal of knowledge enrichment is to make this implicit information explicitly available. CATCH aims to develop knowledge and to demonstrate its applicability in automated knowledge enrichment tools. One group of tools aims to support experts. Another group of tools enables fully automated analyses.
There are two dimensions in these two groups of tools. First, tools can be used to assist experts, or they can perform fully automatically. Second, tools can follow existing annotation schemes, or they can discover new structures within, and relations between objects. Knowledge enrichment can be applied to any of the media types which are covered by CATCH: text, images, handwritten documents, archaeological objects, etc.

Both groups of tools aim to alleviate the following problems occurring in the daily work of collection managers, and in the quality of many existing databases, respectively.

- Cultural heritage experts (collection managers and researchers) have used and developed content annotation schemes and classifications, laid down in thesauri, reference lists, topic maps. Their ability to apply these schemes and classifications to new data is only limited by time and scale. Knowledge enrichment techniques can alleviate the time and scale bottlenecks by adding machine power to manpower; by emulating how experts annotate data. After they have learned to emulate experts by examples, they can start to annotate (classify, analyse, relate) very large amounts of new data themselves, in a fraction of the time.
- Existing databases of objects, partially or inconsistently marked up with legacy classification systems can be automatically made more consistent with knowledge enrichment techniques. As far as they are partially or largely unannotated, disorganized, and unlinked, they can be automatically annotated, organized and linked semantically.

Research topics
The leading question is: How can we arrive at the automatic enrichment of cultural heritage data? We know that the current state of affairs asks for (1) tools to support experts in their manual enrichment work, to alleviate time and scale bottlenecks, and (2) tools for automatic data enrichment, particularly for making existing data cleaner and more consistent, and for discovering new structures and relations in data.
The research agenda that follows from these desiderata starts with the development of methods and software tools that can assist experts in their manual work, allowing them to enrich more data in less time. Such tools should be able to emulate experts' annotations, and suggest annotations of new data at such a high level of precision that experts only need to correct these suggestions occasionally. As a second step, the agenda should list the development of tools that operate in domains that demand even more automation; either because no initial annotation scheme is available (the data is still "raw") and an annotation needs to be bootstrapped from data, or because the annotation needs to be performed automatically, either due to the unavailability of experts or as an initial phase in exploring "raw" data.

This agenda calls for the use and development of methods for automatic knowledge generation in data (a broad field encompassing methods from machine learning, statistical learning, and data mining). Knowledge generation from data is typically needed in situations such as the one central to CATCH, where a digitisation effort has produced (potentially large-scale) databases of unanalysed data, and experts (collection managers) are eager to explore and analyse this data as effectively as possible in as little time as possible. Alternatively, the data is already annotated, or is receiving new annotations through a metadata project (as also present in CATCH), and knowledge enrichment is used to learn this annotation and apply it to yet unanalysed data.

This research is intrinsically empirical; the methods to be developed are based on empirical data, and the function they have can and must be judged and evaluated in terms of measurable improvements in accuracy and speed, both by objective quantitative evaluation and by the collection managers that use the methods.

Background

To understand the research question and the research topics more in depth, we provide some background. Table 1 shows four types of knowledge enrichment we distinguished.

	Expert support	Automatic enrichment
Existing annotation systems	A Expert support, based on existing annotation schemes Supporting experts in the annotation of objects in databases according to an existing annotation scheme, in a software annotation environment that is able to make accurate suggestions. Keywords: semi-automatic annotation, domain knowledge, existing ontologies, semantic web	B Automatic enrichment, based on existing annotation schemes Automatic annotation of unannotated objects, and automatic cleanup of incorrectly annotated objects. Allows to do what under quadrant A could not have been done in human time. Keywords: data mining, text mining, automatic classification, machine learning
Automatic discovery of structure	C Expert support, automatic discovery of structure Confronting experts with statistically salient patterns and structures within and between objects, visualising associations, suggesting new structures. Keywords: exploratory data analysis, data mining, statistical analysis.	D Automatic enrichment, automatic discovery of structure Discovering structures within and between objects, and exporting these discoveries to ontologies, associative networks, and clustering. Keywords: knowledge generation from data, self-organization, clustering

Table 1: Four types of knowledge enrichment.

The "A" quadrant represents tools for the direct support of experts in the manual annotation of objects in databases. Precious time can be saved when intelligent software makes accurate suggestions to the annotator, who then only invests time when the suggestion is incorrect. Even more precious time can be saved when the same intelligent software running in the background makes preselections of especially salient objects that need to be annotated first.

The "B" quadrant takes over from the "A"-quadrant tools when the scale of the data cannot be tackled by the available human expert time. "B"-quadrant tools automatically annotate large amounts of data, and check for inconsistencies and noise in existing annotated databases. They will not do this flawlessly, but well enough that the automatically annotated data becomes largely searchable and retrievable, where before it was not.

The "C" quadrant is the mirror of the "A" quadrant, except that experts are not helped with annotation, but rather confronted with new patterns and relations that may deserve a new annotation symbol or level. A likely example is a new level of annotation which links pairs of objects to each other on grounds of some significant co-occurrence of the two, that thus far was not acknowledged by any level of annotation.

The "D" quadrant combines "B" and "C" - it operates autonomously in data to discover any grouping of objects that might be of interest, on such large amounts of data that a manual inspection of the process would not be feasible, except at the very end of the automatic knowledge discovery process.

2.1.3 Theme 3: Personalisation through presentation

Situation in cultural heritage

Most of the services that are currently available have predefined presentations. The institutions determine the ways a user may view objects and their metadata. Information technology offers many new options for personalisation of the presentation, but these are hardly used at all. The reason is straightforward: there are actually no easy-to-use tools in that respect. More research into human-computer interaction and user modelling is needed to specify such tools. A clear instance is the need for better navigation through digital collections. The amount of objects from cultural institutions run in the millions, if not billions when considered on a global scale. User modelling is considered as an attractive option for navigating more quickly, easily and efficiently across digital collections or objects. By automatic analysis of the user's search behaviour and by offering the facility to create personal contexts, it is expected that users can benefit more from such information services than via direct search-and-retrieval actions.

Research topics

The leading research question is: How can we develop methods and tools for generating presentations of cultural-heritage objects that are related in a semantic way? This work also includes (1) user-modelling issues, e.g., how can user groups be related to presentation styles? and (2) user-control issues, e.g., how can the user control the presentation style? More specifically, we list the following three research questions.

- Is it possible adequately to reduce the user's effort when expressing the ambitious information need that the system must take into account besides many other elements?
- Is it possible to construct a tool that composes an agreed-upon ontology in order to determine the meaning of terms in the user's questions and in the information sources?
- To what extent is it possible to find an "optimal" mix of (1) proactive behaviour that is based solely on the user's known interests and (2) selection of information based on other users' interests or the importance of certain (unrequested) information?

For the research involved two observations are important.

- The availability of a syntactically (XML-based) and semantically (RDF/OWL based) integrated metadata opens new avenues for presentation and personalization.
- By using semantic relations such as "period" and "style" it becomes possible to generate tailor-made presentations for groups or individuals.

Background

To provide an appropriate insight into the complexity of the three research questions we add some details about context and depth of the investigations. In research question 1, the "many other elements" include a user model containing the interests, goals, background and knowledge of the user, contextual information such as the physical location of the user and perhaps also his/her orientation, the time of day, the device and network he/she is using to interact with the system. Presently research is carried out on adapting the selection and presentation of information to a user based on one type of information about that user (either knowledge, interest, or context). This should be complemented by research on adaptation based on all kinds of information about the user in question and his/her context.

For research question 2 it is beneficial to understand that the answer to a question also consists of objects described by semantic metadata, used to determine how these objects relate to one another. This semantic information needs to be combined with descriptive metadata in order to generate a hypermedia (Web) structure that can be viewed using a "browser". While currently it is possible to generate such presentations based on one set of metadata, the combination of different types of metadata has to be investigated in order to generate the most appropriate presentation for each individual user.

Research question 3 looks somewhat further into the future: systems can be made to become proactive, selecting and presenting information that matches the user's interests and needs without the user having to express that need through a question. The automatic provision of information on a person, e.g., architect Max Weber, when dealing with housing of multicultural groups in Amsterdam, is a good example of proactive behaviour. A mix of active and proactive behaviour is needed in order to prevent an agent from becoming boring because an agent will never surprise the user with interesting but unexpected information.

For the research theme personalisation the CATCH programme aims at acquiring new knowledge in three subdomains: (1) selection of information, (2) automatic generation of presentations, and (3) adaptation or personalisation.

Selection of information. The challenge here is to answer incomplete information requests from users with an accuracy that is comparable with or even better than the database-query accuracy. Four techniques have to be combined into heuristic evaluation tools to achieve this goal. The techniques are: (1) information retrieval techniques based on (potential) natural language understanding of textual contents, (2) information retrieval techniques based on metadata using ontologies, (3) selection of objects based on descriptive metadata, and (4) database integration methods.

Automatic generation of presentations. The challenge is to "combine" selected information objects of different media types. Perhaps having different types of navigational or semantic relationships and combining them into a single virtual hypermedia (Web) presentation is the most difficult part. In that case it is necessary to adapt the result to the device and network capabilities of the user's environment. This requires a careful (automatic) selection of the use of the "dimensions" layout, time, and navigation.

Adaptation or personalisation. The results of almost any possible information request are too large to be presented to and browsed through by a user. Hence, an environment must be designed that derives additional specifications of the information or objects to be selected from past user behaviour. In order to improve this process, and especially its initial stages, users need to be clustered in groups (with similar interests, background, expertise, etc.). Finding scalable algorithms for grouping is an additional research issue here.

2.2 Implementation Strategy

The implementation strategy has two branches: the practical implementation and the structural implementation. The practical implementation focuses on the character of the project: demand pull. Hence in 2.2.1 we discuss "tools for the back office" and in 2.2.2 we deal with the composition of the research teams and their collaborations. The structural implementation emphasizes the design principles to be valid for all cultural heritage institutions and to be followed by all research teams (in 2.2.3). In 2.2.4 attention is paid to the connectedness of the knowledge suppliers (the cultural heritage), the researchers, and the end users by introducing two integrators in which the software and tools have to be implemented.

2.2.1 Tools for the "back office"

The potential users of the results of CATCH fall into two categories.

1. The collection managers of the cultural heritage institutes.
2. The end users of the services provided by the cultural heritage institutes.

The two categories have their own demands. The first group is located in the "back office". Here preparations are made for the services and products (such as exhibitions, catalogues, and websites) which will be presented to the end users: the people who are the rationale for the very existence of the cultural heritage institutions. Within the category of end users we distinguish four groups.

a. Research: scientific staff from disciplines like History, History of Art, Archaeology, Cultural Studies, Linguistics, etc.
b. Education: teachers at universities, high schools, Art Academies.
c. Media: journalists, publishers, editors, marketeers of cultural heritage institutions.
d. Entertainment and edutainment: the general public.

The CATCH research focuses on the development of tools and methods for the collection managers of the cultural heritage institutes (category 1 users) that will enable them to do more in less time and with higher quality. This speeding up of back office processes is needed for at least three reasons: (1) the rapidly growing amount of digitised heritage, (2) the existing amount of heritage that is still waiting to be processed and (3) the ever fastening changes in public demand (category 2 users). Cultural heritage institutes have to adapt to these changes or they will become obsolete. Information technology can provide tools to support the back office in their endeavour to enhance the interaction between the end users and their cultural heritage. It is the ambition of CATCH to develop new knowledge and demonstrate its applicability in a number of tools suitable for use in wide ranges of cultural heritages institutes.

Within the category of end users CATCH pays special attention to group (a): scientific staff from disciplines like History, History of Art, Archaeology, Cultural Studies, Linguistics etc.

2.2.2 Composition of the Research Teams

Essential for the rationale underlying the CATCH programme, the temporary researchers and programmers financed by CATCH will be employed by the universities[3] but will have their daily work within the cultural heritage institutes. By physically locating the researchers in the environment where the fruits of their research will be used, CATCH aims at supporting a vivid interaction between the researchers and the prospective users. The idea is that the principal investigator remains responsible for the quality of the research being done, and that the director of the hosting cultural heritage institute has control over the daily routine. Rights and duties of all parties involved are laid down in a guest researcher agreement.

The CATCH research teams will consist of:

- CATCH-funded temporary researchers (PhD students, postdocs), employed by the universities.
- Senior research staff employed by universities.
- Senior staff employed by cultural heritage institutions (researchers and/or collection managers or others with relevant expertise).
- CATCH-funded temporary scientific programmers, employed by the universities.
- Programmers employed by the cultural heritage institutions.

Each team is a mix of persons from each of these five categories. The CATCH principles of interaction and co-operation is also manifest in the composition of the research teams. The PhD-students, postdocs and programmers financed by CATCH are embedded in a team consisting of both senior researchers from one or more universities and senior staff from the cultural heritage institute acting as host. The teams are jointly headed by the principal

[3] *In this programme text "universities" is used as a shorthand for "universities, Telematics Institute and Max Planck Institute for Psycholinguistics".*

investigator from the university and one of the senior staff members of the hosting cultural heritage institute.

The programme starts with the formation of six teams. For each team, CATCH funds one PhD student (four years), one postdoc (three years) and one scientific programmer (four years). The six teams will each execute a *core project*, which together constitute the foundation for the research programme. The research details of the core projects are given in Appendix I. Table 2 gives an overview of the universities and cultural heritage institutions involved in the core projects. The second and third column mention the principal investigator and staff member cultural heritage who are jointly responsible for the execution of the project. The fourth column mentions the universities which will employ the researchers and programmers. The last column mentions the cultural heritage institutions in which the researchers and programmers will actually do their work.

	Principal investigator & university	**Staff member cultural heritage**	**Researchers & Programmers**	
			University	**CH Institution**
Theme 1: Semantic interoperability through metadata				
Project 1.1: STITCH	Van Harmelen, VU	Matthezing, KB	1 PhD VU 1 Postdoc MPI 1 Progr. VU	KB
Project 1.2: CHOICE	Veenstra, TI	Oomen, B&G	1 PhD TI 1 Postdoc VU 1 Progr. MPI	B&G
Theme 2: Knowledge enrichment through automated analyses				
Project 2.1: RICH	Postma, UM	Lange, ROB	1 PhD UM 1 Postdoc UM 1 Progr. UM	ROB
Project 2.2: SCRATCH	Schomaker, RUG	Jager, NA	1 PhD RUG 1 Postdoc RUG 1 Progr. RUG	NA
Project 2.3: MITCH	Van den Bosch, UvT	Houtgraaf, Naturalis	1 PhD UvT 1 Postdoc UvT 1 Progr. UvT	Naturalis
Theme 3 : Personalisation through presentation				
Project 3.1: CHIP	De Bra, TUE	Sigmond, RM	1 PhD TUE 1 Postdoc TI 1 Progr. TUE	RM

Table 2: Distribution of core projects over themes, universities and CH institutions

Universities:

MPI	= Max-Planck-Institut für Psycholinguistik, Nijmegen
RUG	= Rijksuniversiteit Groningen
TI	= Telematica Instituut, Enschede
TUE	= Technische Universiteit Eindhoven
VU	= Vrije Universiteit, Amsterdam
UM	= Universiteit Maastricht
UvT	= Universiteit van Tilburg

Cultural Heritage Institutions:

B&G	= Nederlands Instituut voor Beeld en Geluid, Hilversum
KB	= Koninklijke Bibliotheek, Den Haag
NA	= Nationaal Archief, Den Haag
Naturalis	= Nationaal Natuurhistorisch Museum Naturalis, Leiden
RM	= Rijksmuseum, Amsterdam
ROB	= Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort

During its lifetime, CATCH will be able to fund a total of 17 of research teams, i.e., 34 temporary researchers and 17 programmers[4]. The 11 remaining teams will be selected in competition on the basis of research plans. All Dutch universities can enter the competition, which will be organised by NWO and obey the usual NWO rules and regulations for research programmes like this.

2.2.3 Design Principles

CATCH focuses on knowledge-based access of the cultural heritage (sources, resources, and knowledge). IT provides tools to facilitate access. Three themes are formulated to guide the research and development of tools: semantic interoperability, knowledge enrichment, and personalisation. Moreover, strategy and organisation determine the constraints that the projects must meet. The software developed will have the character of open-source software.

The CATCH programme should start with determining a standard measure, i.e., an inventory of what is available on (say) November 1, 2004. This will be done in two respects. All PhD students and postdocs will start their project with a 'warming-up period' of two month to get acquainted with the state of affairs in their hosting cultural heritage institution. During this period they become aware of the problems the cultural heritage institution encounters in their IT-operations. The focus is, of course, on problems related to the research project to be executed. It is very important that during this period the researchers (and their supervisors) get to know the organizational structure of the hosting institution and the people outside the research team (often support staff) who can in some stage contribute to the progress of the actual research effort. One practical way of doing this, is by tackling a small practical IT-problem. This will benefit both the researchers (who will get a crash course about the hosting institution) and the hosting institution (who will have one of their small IT-problems solved).

The warming-up for all programmers consists of making an inventory of the existing and emerging (software) standards relevant to their hosting institutions. In a later stage the inventory can be broadened to requirements for standardisation accepted in the cultural sector.[5] The inventory is very important, since the group of programmers will be responsible for implementing the interoperability results obtained by the researchers. The inventory can help in further focusing the research effort as the programme progresses.

[4] *The exact mix of personnel is to be determined during the execution of the programme.*

[5] *DEN is the organisation that guards these standards. DEN is in contact with the Netherlands Standardization Institute NEN. Both organisations investigate forms of co-operation in the field of digital cultural heritage.*

Six organisational principles will lead to a uniform development of rules for projects within CATCH (see 3.2). Below we provide the notions of the guidelines which will be set up in the first phase of the project. CATCH design principles are as follows.

- Distributed systems
- Extreme modularity
- Open standards
- Web enabled systems
- Interoperability
- Use of adaptive IT Techniques
- Digital durability

2.2.4 Integrators

To optimise the success factor and to assure the interoperability the software has to be implemented into at least two integrators: (1) The Memory of the Netherlands (a large database and website about digitised cultural objects maintained and developed by the Koninklijke Bibliotheek) and (2) a museum environment (e.g. the Rijksmuseum). Of course, the application must also be able to work with systems in use in the host cultural heritage institution. No software will be accepted that only runs in just one environment. Knowledge and software must contribute to the integration and interoperability of collections of participating cultural heritage institutions as well as non-participating institutions. The programme committee's second task is to see to it that all the software developed in each of the projects is embedded in at least one of the CATCH integrators.

The CATCH programme is structured according to three themes. All cultural-heritage institutes participating in the CATCH programme will be involved in the research lines in the first phase of the project. Figure 2 illustrates the general structure of the programme. The integrators form the centre of the programme, the testbeds where all techniques and methods come together. Going from the bottom to the top of the diagram, we observe the following stages. The cultural heritage institutions (depicted at the bottom of the diagram in Figure 1) digitise their heritage objects. Durable storage and knowledge enrichment techniques operate on the digitised objects.

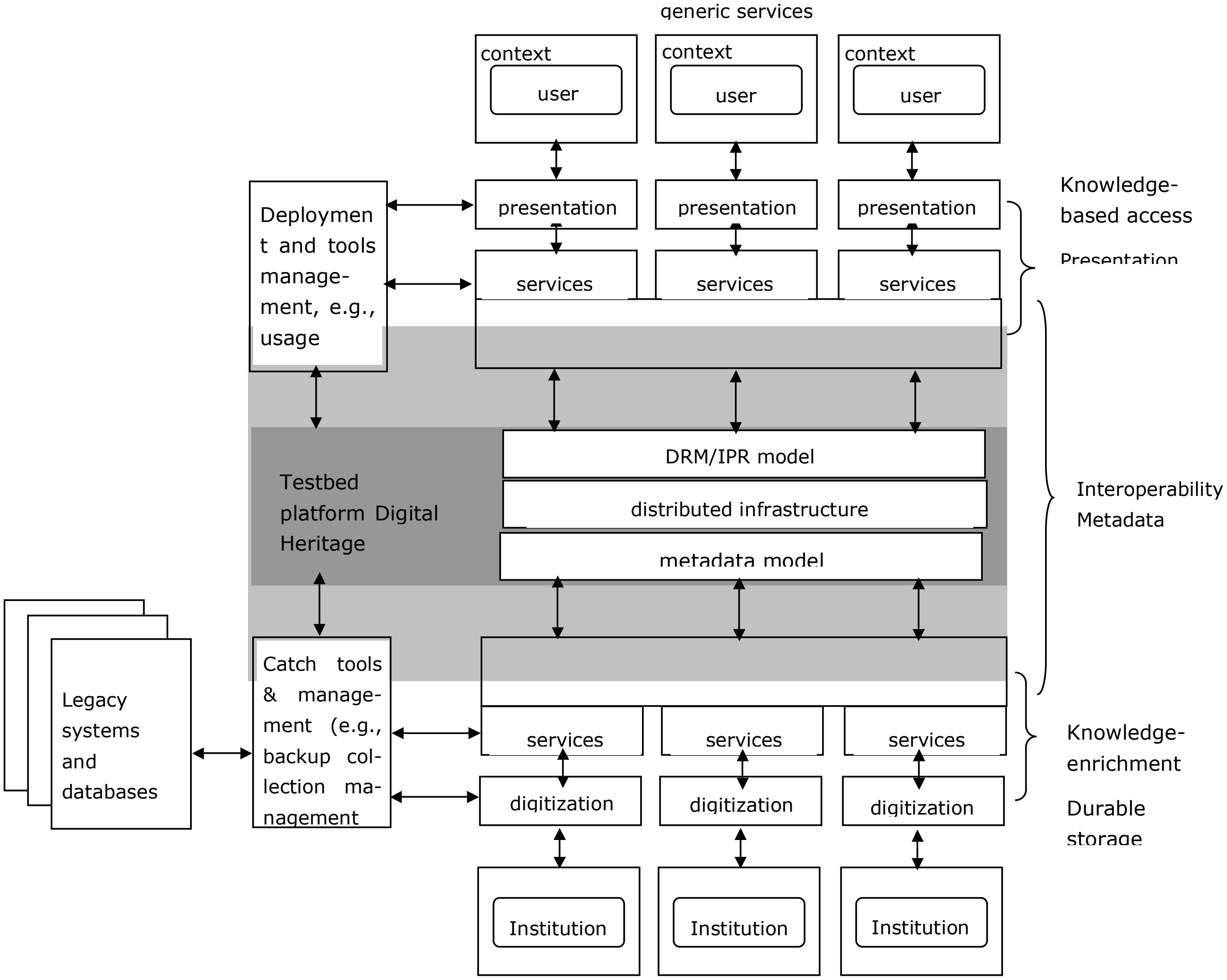


Figure 1: Schematic illustration of the integrators and its relations to the CATCH research themes.

The results generated by, for instance, enrichment techniques lead to novel metadata. Within the integrator (the shaded area in the diagram), a metadata model is specified that prescribes the format of the newly created metadata. In addition, the integrator realises a distributed infrastructure in which the research line of interoperability plays a main role. At the user side (depicted at the top of the diagram), the research lines of theme 3 personalisation will enhance the accessibility for the user. Thus the integrators play a pivotal role in the CATCH programme: all research themes come together within or at the boundaries of the integrator.

3. SUPPORT STRATEGY

The CATCH programme is a demand-pull programme, with the aim to perform excellent research and produce tools and software that are valuable to the cultural heritage institutions. However, achieving such a twofold aim is not sufficient to boast in the near future on a successful project. Therefore, a support strategy has to be developed, in the form of a support programme with two aims.

1. To facilitate the transfer of knowledge and tools (a) within the programme and (b) to all other parties interested in the CATCH results.
2. To build and establish a structure which guarantees continuity for the results (in particular the tools, the software, and the knowledge) of the programme.

The support programme is run by the Programme Management Bureau (see section 4.4).

3.1 Transfer of knowledge and tools

In the programme’s first year the Programme Committee will formulate and implement a specific plan for knowledge transfer. The costs of this plan will amount to approximately 10% of the research budget. The following seven items list the initiatives to be implemented.

Publications: The results of the fundamental strategic research will be published in the usual scientific media (doctoral theses, articles in journals, contributions to conferences and workshops).

Demonstrators: Researchers will be stimulated to develop demonstrators showing the potential of research results which can make an important contribution to the knowledge transfer.

Annual Seminars: Every year the CATCH and MultimediaN Programme Committees will jointly organize a seminar, the Dutch Multimedia Event. Furthermore, other seminars may be organised that focus on the Dutch researchers and cultural heritage experts active in fields closely related to the programme. Members of the International Scientific Advisory Board will also be invited to attend. Although these seminars will primarily focus on the Dutch experts, the organisation will invite a number of prominent foreign researchers who will be asked to comment on the status of research in the CATCH programme.

Workshops: Two international workshops will be organized: one after two and a half years and one at the end of the programme. The topics will be selected from the three themes. It is assumed that approximately 100 people will participate in these workshops; the majority of whom will be from abroad.

The workshops will in particular play a role in the programme’s evaluation. To this end, the workshop halfway through the programme can be made to have consequences for the planning of the second half of the programme.

User group: User Groups will be formed in order to guarantee the transfer of knowledge to cultural heritage experts, the business community and society in general. At least three groups will be formed, one for each of the three research themes. Each User Group consists of representatives from interested industrial companies and institutions with a background that enables them to provide substantive feedback on the progress, course and results of the research. Special user seminars will also be organised in consultation with the cultural heritage and business community.

Patents: Patent applications are an important form of knowledge transfer. The CATCH programme will strive to develop patentable knowledge. Project partners will lay down agreements with regard to patents and licenses. STW will assist the possible exploitation of patents.

Website: The programme will maintain a website which will be used to provide companies, institutions and the popular scientific press access to the results of the research. The researchers in the programme will be stimulated and - where necessary - supported so that they can present the results of their research in a way which makes it accessible to outsiders. Furthermore, there will be a members only section on the website which is only accessible to researchers immediately involved in the programme.

Moreover, the Programme Committee will link the programme to initiatives like "Boulevard van het actuele verleden" ("Boulevard of the current past"), which seek to create a historical "experience" for the general public. The aim of "Boulevard" is to submerge visitors in a virtual world, recreating an historical past. The Programme Committee will explore if and in what way CATCH research can contribute to initiatives like "Boulevard".

3.2 Continuity

Initially, knowledge transfer will be promoted by (a) the participation of cultural heritage institutes and knowledge institutions in the Programme Committee who control the research and (b) by the joint participation of cultural heritage experts and academic researchers in the programme projects. More specifically, in the individual CATCH projects the researchers and programmers will be hosted by cultural heritage institutes, i.e., they will actually perform a considerable part of their research within the environment of the cultural heritage institutes thus allowing for optimal knowledge transfer opportunities.

There are six organisational principles imposed by the CATCH programme that hold for all participants.

- The Programme Committee will ensure that the IPR to the software and tools developed within the CATCH programme will be properly secured.
- Tools and software developed within the CATCH programme must be centrally registered after completion of the project (during the development they will be provisionally registered). The Programme Committee has already established preliminary discussions with SURF about the support, maintenance and availability of the tools and software that will be developed within the CATCH programme (c.f. DARE repositories[6]).
- Tools and software are freely available and usable for the partners. Moreover, they will also be made available for cultural heritage institutions which do not directly participate in CATCH. However, these institutions should register their use of the tools at the administration controlling the software and tools.
- Cultural heritage institutions may elaborate on the software obtained. However, they have the duty to supply their results for free to the organisation serving as a clearing house for the CATCH programme results.

[6] *SURF DARE repositories: http://www.darenet.nl/en/toon*

- Commercial partners have the right to exploit the software developed in the projects in which they participated. However, they may not do so exclusively.
- Commercially interested partners from outside the projects can have such rights granted after explicit permission of the IPR-owner of the CATCH results, which can impose constraints or financial obligations.

The results of the research projects will in most cases partly consist of newly developed software and algorithms. Portability of these results will be stimulated by the design principles given in 2.2.3.

The Steering Committee and the Programme Committee will ensure the continuity of the programme efforts by making specific arrangements with SURF and DEN with respect to the continued availability and maintenance of the programme results with respect to software and algorithms after the project has ended.

4. PROGRAMME MANAGEMENT AND BUDGET

This section contains an overview of tasks and responsibilities of three committees and the Programme Management Bureau. Furthermore, a global overview is given on the budget.

4.1 Steering Committee

The Steering Committee of the CATCH programme will be formed by the members of the Council for Physical Sciences supplemented by at least one representative of the Council for Humanities and a representative of the cultural heritage institutes. A SURF representative will also be invited sit in as an advisor. If other parties decide to contribute financially to the CATCH programme, the composition of the Steering Committee may be extended. The Steering Committee meets twice a year, or more often if necessary.

The tasks and responsibilities of the Steering Committee (SC) are as follows.

- The SC supervises the Programme Committee (PC) in the execution of the research programme with regard to progress and cohesion.
- At least once a year the SC reports to the financing bodies of the programme about the progress of the programme and its financial situation.
- The SC formally appoints the members of the Programme Committee.
- The SC every year has to approve the PC's proposal for the budget.
- The SC makes the formal granting decisions on the basis of a PC proposal.
- The SC ensures that specific actions are taken to ensure the continued availability and maintenance of the programme results.

4.2 Programme Committee

The Programme Committee (PC) is appointed by the Steering Committee. The PC will consist of maximally 12 persons, who will be appointed on the basis of their expertise related to the CATCH programme. The Programme Committee will consist of:

- the two programme leaders
- the leaders of the three research themes: per theme one computer science and one CE representative
- some representatives of related programmes.

The directors of the NWO Councils for Physical Sciences and Humanities will have a standing invitation for the meetings of the Programme Committee.

The tasks and responsibilities of the Programme Committee are as follows.

- The PC determines and monitors the course of the research programme.
- Within six months after the start of the programme, the PC will submit to the SC a list of success criteria which are to be used in evaluating the programme.
- Before the end of the first programme year, the PC will formulate a specified plan for knowledge transfer.
- The PC formulates Calls for Proposals, appropriate research themes and assessment criteria.
- Each year the PC reports to the SC about the progress of the research programme, its budgetary situation and its plans for the next years.

- The PC is responsible for organising a midterm evaluation and a final evaluation.
- At least three times a year the PC will organise a meeting at which all the researchers involved in the programme will present their results and their plans for future research. Foreign experts can be involved in these seminars.

The programme leaders, the programme manager and the directors of the Council for the Physical Sciences and the Council for the Humanities form an Executive Committee, which will be responsible for handling the day-to-day affairs.

4.3 International Scientific Advisory Board

The Programme Committee and Steering Committee will be assisted by an International Scientific Advisory Board (ISAB), consisting of internationally respected experts in the field of information science and the application of these techniques on cultural heritage data, and specialists from cultural heritage institutes with expertise in computer science. The ISAB functions as an external assessor of the six core projects that will form the basis of the CATCH programme. These projects can only start after approval from the ISAB. Moreover, the ISAB will review and prioritize the full proposals submitted in the competitions (section 4.7). Annually, the SC seeks the ISAB´s advice on the quality and the direction of the CATCH research seen in international perspective. The ISAB will also be involved in the midterm and final evaluation of the project. Finally, the members of this board will be invited to attend the CATCH workshops and can be consulted as advisors for those involved in the CATCH project.

4.4. User Groups

As was already mentioned User groups will be formed in order to guarantee the transfer of knowledge to cultural heritage experts, the business community and society in general. At least three groups will be formed, one for each of the three research themes. Each User Group consists of representatives from interested industrial companies and institutions with a background that enables them to provide substantive feedback on the progress, course and results of the research. Special user seminars will also be organised in consultation with the cultural heritage and business community. The chairman of each User Group will be part of the Programme Committee. These User Groups will also be actively involved in determining the programme´s direction and in evaluating the progress of the individual projects and the programme as a whole.

4.5 Programme Management Bureau

The SC, PC, and ISAB will be supported by a Programme Management Bureau (PMB) which will be hosted by NWO. The CATCH PMB consists of a programme officer and his/her staff.
The PMB costs will be covered by the programme budget.

The tasks and responsibilities of the Programme Management Bureau are as follows.

- The PMB supports the SC, the PC and ISAB and prepares their meetings.
- The PMB is responsible for the day-to-day scientific managerial and financial administrative affairs of the programme.

- The PMB organises the calls for proposals.
- The PMB monitors the progress of the programme projects and formulates the yearly progress reports.
- The PMB stimulates the coherence and knowledge transfer within the programme.
- The PMB promotes the dissemination of the programme results.
- The PMB takes care of the practical organisation of programme workshops and evaluations.

4.6 Committee of Recommendation

The Cultural and Industrial Advisory Board will consist of a number of persons with an influential cultural or industrial position in the Netherlands who have agreed to function as ambassadors for the CATCH programme.

4.7 Budget

The total budget of the programme is estimated at M€ 15,3, of which M€ 12,5 will be made available as subsidies and M€ 2,8 will be contributed in kind by the participating cultural heritage institutions. The programme starts with M€ 6,0 in subsidies, committed by the NWO Councils for Physical Sciences and Humanities. The remaining M€ 6,5 in subsidies have been reserved by NWO (M€ 5,0) and the Ministry of Education, Culture and Science (M€ 1,5), but their definitive commitment to the programme depends amongst others on the progress the programme makes.

The in kind contribution of the cultural heritage institutions will be 25% of the subsidies provided by NWO. The contributions will be realised through the participation in the CATCH research teams of researchers, programmers and other staff employed by the cultural heritage institutions (cf. section 2.2.2), and through the participation of representatives of the cultural heritage institutions in CATCH's governing bodies.

Section 5.1.2 describes developments within the Royal Netherlands Academy of Arts and Sciences (KNAW) regarding a programme e-Science for humanities and social sciences. If granted, the programme is expected to have a budget of M€ 4,5. Although the programme will not be part of CATCH in the strict sense, there are clearly related issues in both programmes. Coordination and linkage is secured by the participation of Peter Doorn of the KNAW in the (preparatory) CATCH Programme Committee. If the KNAW programme is granted, it will contribute to the joint national effort with respect to the accessibility of digitised Dutch cultural heritage. In the table below the KNAW programme is added provisionally.

(Amounts in M€)	**Phase 1**	**Phase 2**	**Total**
NWO Physical Sciences	5,0	2,5	7,5
NWO Humanities	1,0		1,0
NWO General Board		2,5	2,5
Ministry of Education, Culture and Science		1,5	1,5
Total Subsidies	6,0	6,5	12,5
Contribution cultural heritage institutions	1,5	1,3	2,8
Total CATCH programme	7,5	7,8	15,3
KNAW e-Science for humanities and social sciences	PM	PM	(4,5)

The budget is available for the execution of the three CATCH strategies: research, implementation and support. As described in chapter 2 and 3, these strategies are closely intertwined. The preliminary distribution of the budget over the three strategies is depicted in figure 2.

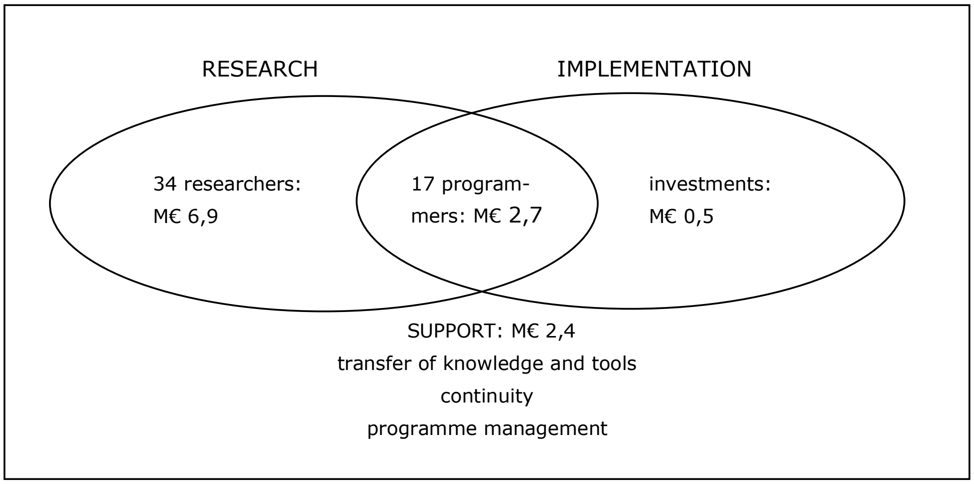


Figure 2: Preliminary distribution of the budget over the three CATCH strategies

Assuming an average project budget of k€ 565, a total 17 of projects can be funded. The subsidy allows for the payment of the wages for one PhD student, one postdoc for three years and one programmer for four years. Furthermore, within each project budget k€ 24 is available for the purchase of small computing equipment and software, on top of the usual bench fee of k€ 5 for each PhD student and postdoc.

The programme starts with six core projects. The eleven remaining teams will be selected in competition on the basis of research plans. All Dutch universities can enter the competition, which will be organised by NWO and obey the usual NWO rules and regulations for research programmes like this. The CATCH competitions will be part of the annual competition for NWO computer science programmes (call for proposals in November, deadline for submission in February, decision for acceptance/rejection in July).

Assuming a more or less even distribution of the research budget over the three research themes, the relation between core projects and projects to be granted in competition is:

(Amounts in k€)	core projects[7]		competition		total projects	
	no.	budget	no.	budget	no.	budget
Theme 1	2	1.130	4	2.260	6	3.390
Theme 2	3	1.695	3	1.695	6	3.390
Theme 3	1	565	4	2.260	5	2.825
Total subsidy	6	3.390	11	6.215	17	9.605
Contribution CH[8]		848		1.554		2.402
Total research		4.238		7.769		12.007

Theme 1 = Semantic interoperability through metadata

Theme 2 = Knowledge enrichment through automated analyses

Theme 3 = Personalisation through presentation

For all budget figures holds that the actual distribution can be adjusted by the Programme Committee and the Steering Committee depending on the development of the programme or advise of the International Scientific Advisory Board.

[7] *In fact, the budget for the core projects is k€ 3.210 (and thus the budget for the other projects k€ 6.395), since the wages for the researchers and programmers are lower in 2004 than they will be in later years. For ease of presentation, the* average *project budget of k€ 565 has been used in this table.*

[8] *On top of the k€ 2.400 mentioned in this table, the cultural heritage institution will contribute k€ 400 through the participation of their representatives in the Programme Committee, Steeering Committee and International Scientific Advisory Board.*

5. NATIONAL AND INTERNATIONAL CONTEXT

Digital access to cultural heritage for the general public as well as education and humanities research has become an important policy area since the second half of the 1990s. At the G7 Conference on the Information Society in 1995, the potential offered by Information Technologies for "Multimedia Access to World Cultural Heritage" was officially recognized. Since then, "digital heritage" and "e-culture" took important positions on the political agenda of the information society in many countries and international organizations. It is hardly possible to sum up the programmes and projects that were set up in the past decade in the field of digital culture. Nevertheless, this section aims to give a broad overview of the context in which the CATCH programme can be placed, both nationally (in 5.1) and internationally (in 5.2).

5.1 National context

In 1997 the Royal Netherlands Academy of Arts and Sciences (KNAW) published a report calling for enhanced digital access to cultural heritage information and improved ICT for humanities research.[9] In 1998 the report *Alles uit de Kast*[10] outlined the contours of a national investment programme for establishing a digital infrastructure for cultural heritage. This was followed by a plan by NWO to create a virtual digital research library for the humanities.[11] In the beginning of 2002 the *eCultuurnota*[12] appeared. The report sketched the outline of a digital infrastructure for the cultural domain. In particular, the report identified the need for enhanced accessibility of cultural sources and the possibility of reusing cultural material. In May 2002 the governmental letter *Digitalisering van het Cultureel erfgoed*[13] appeared. The letter described in more detail how the digitalisation of the cultural heritage should come about.

Meanwhile, in 2000, the Ministry of Economic Affairs had published a report called *Concurreren met ICT-Competenties, Kennis en Innovatie voor De Digitale Delta*[14] emphasizing the importance of enlarging ICT competence in the Netherlands. In 2001 the taskforce "ICT-en-kennis" (the Le Pair Committee) issued the report titled *Samen, Strategischer en Sterker*[15] recommending the exploitation of scientific expertise in the multimedia sector to develop new application areas.

[9] De computer en het alfaonderzoek*. Advies van de Commissie Geesteswetenschappen over de toepassing van de informatietechnologie bij het onderzoek op het gebied van de geesteswetenschappen, voorbereid door de Subcommissie Informatietechnologie Alfaonderzoek (1997) KNAW.*

[10] *Alles uit de Kast – Op weg naar een nationaal investeringsprogramma digitale infrastructuur cultureel erfgoed (1998). Wetenschappelijk Technische Raad SURF.*

[11] *Een Digitale Bibliotheek voor de Geesteswetenschappen. Aanzet voor een programma voor investering in een landelijke kennisinfrastructuur voor geesteswetenschappen en cultuur (december 1999). NWO-Gebiedsbestuur Geesteswetenschappen.*

[12] eCultuur in Beeld*, letter of the Dutch Parliamentary Undersecretary van der Ploeg to the Tweede Kamer der Staten Generaal on April 22 2002 (Kenmerk MLB/M/2002.14.192).*

[13] Digitalisering van het cultureel erfgoed*, letter of the Dutch Parliamentary Undersecretary van der Ploeg to the Tweede Kamer der Staten Generaal on May 27 2002 (Kenmerk DCE/02/18765).*

[14] Concurreren met ICT-Competenties.Kennis en Innovatie voor De Digitale Delta*, report of the Dutch Minister of Economic Affairs A. Jorritsma-Lebbink and Minister of Education Drs. L.M.L.H.A. Hermans, Onderwijs, Cultuur en Wetenschappen April 2000.*

[15] Samen, strategischer en sterker*, final report of the Taskforce ICT-en-kennis (Committee Le Pair). April 2001.*

The growing policy relevance of innovative digital techniques for the domain of cultural heritage and the humanities is an international phenomenon. Research into virtual libraries and museums, digital longevity of archival sources, techniques of digitization and access to cultural content is taking place in many countries by researchers from computer and information science, humanities computing and the heritage sector itself.

The umbrella organisations for the sciences and humanities in the Netherlands (KNAW, NWO and SURF; a brief overview of their activities is given below in 5.1.1, 5.1.2, and 5.1.3, respectively) have started to develop new plans to give a strong impetus to the intersection of computing, heritage and humanities.[16] Meanwhile, computer and information science is increasingly aware of the research challenges posed by the cultural domain. In the national research agenda for computer science 2001-2005 (NOAG-i) this domain is present in several themes and programs (e.g., ToKeN 2000, Cognition, Language and Speech Technology). In section 5.1.4 we provide some information on MultiMediaN.

5.1.1 The Royal Netherlands Academy of Arts and Sciences

On the basis of several commission reports regarding the future of the Netherlands Institute for Scientific Information Services (NIWI), the KNAW has decided to start an e-Science programme for the humanities and social sciences.[17] The new program is part of a broader KNAW policy aiming at significant advances in the effective use of ICT in the humanities and social sciences. This new policy includes actions on different levels: principles of open access to research output and data, investments in ICT infrastructure, and the establishment of data archiving networked services (jointly with the Netherlands Research Council NWO). With this new e-science research program, the KNAW seeks to fuel the development of this emerging field in the Netherlands and achieve a leading position internationally.

The KNAW e-science program needs to address a dual mission: (i) to stimulate the development of e-science in the humanities and social sciences, and (ii) to study the effects of e-science on the practice, activity and quality of research in those fields. This mission is to be pursued by an integrated program of cooperative research between the humanities, social sciences and information sciences.

The development of ICT and in particular the Internet, have brought significant changes in three areas: (i) the ever-growing availability of computing power, both in the personal computer and through the emerging GRID technologies linking many computers together;

[16] *NWO with the present Catch plan; the KNAW with a programme on e-Science in the humanities and social sciences, cf.:* Building the KNAW International Research Institute on e-Science Studies in the Humanities and Social Sciences (IRISS) *Committee on a KNAW Research Institute for e-Science (Chair: Prof. dr. ir. Wiebe E. Bijker) (2003) KNAW; SURF has published the report* E-based Humanities and E-humanities on a SURF platform, *by Joost Kircz (2004) SURF.*

[17] KNAW (Commissie van Bemmel), E-wetenschapsonderzoek in het alfa- en gamma-domein, Advies van de tijdelijke commissie Strategie NIWI-KNAW. Koninklijke Nederlandse Akademie van Wetenschappen (Amsterdam, 2002). Commissie Informatiediensten NIWI (voorzitter: dr. N.M.H. van Dijk), *Behouden Toekomst: Een advies met betrekking tot de toekomst van de diensten van het Nederlands Instituut voor Wetenschappelijke Informatiediensten* (Amsterdam, 2003). Committee on a KNAW Research Institute for e-Science (Chair: Prof. dr. ir. Wiebe E. Bijker) *Building the KNAW International Research Institute on e-Science Studies in the Humanities and Social Sciences (IRISS)* (Amsterdam, 2003).

(ii) facilities for communication and collaboration through the internet and applications such as e-mail and the world wide web; (iii) access to digital collections of data, including text, sound and images.

E-science is regarded as the combined use of these advances. Potentially e-science can have a profound influence on research, the questions researchers ask and the way research is carried out. E-science first took off in the natural and life sciences, but interest from the social sciences and humanities is growing rapidly; each of the three areas mentioned above has seen increasing activity. Computers are being widely used, and the growing power has led to new research tools.

On the whole, the development of e-science research practices in the humanities and social sciences appears to be in its early stages. This raises two sorts of questions: (1) To what extent are researchers posing new questions, or are existing questions approached in a different (new) way; are new methods desired and developed, and are new patterns of interaction and cooperation emerging among researchers internationally? and (2): How do researchers organize their electronic environment, what are the problems they encounter and how can these be overcome?

The combination of these two sorts of questions, the one more reflective, the other more practice oriented, necessary to gain new insights into to the new possibilities and pitfalls of e-science, is the essential characteristic for an e-science research programme as envisaged by the Academy.

5.1.2 The Netherlands Organisation for Scientific Research

In 1999, the NWO Research Council for Humanities established a platform to prepare the development of a production line for the Digital Library for the Humanities.[18] It recognized the importance of ICT techniques for providing adequate and broad accessibility to cultural heritage and the possibilities this would create for future research in the humanities. Meanwhile, the Research Council for Physical Sciences launched a cooperation with researchers in the cognition domain. Their project was called ToKeN2000, and one of the major application areas was the cultural heritage sector. As a natural consequence of these two developments, in 2002 both councils joined forces which has led to the present CATCH proposal. In summary, the motivation of NWO reads:

- to stimulate innovative research;
- to encourage cooperation between front-ranked researchers of different disciplines;
- to strengthen ties between researchers, research applications, and society.

5.1.3 SURF

SURF, the higher education and research partnership organisation for network services and information and communications technology in the Netherlands, is active in the field of

[18] *Een Digitale Bibliotheek voor de Geesteswetenschappen. Aanzet voor een programma voor investering in een landelijke kennisinfrastructuur voor geesteswetenschappen en cultuur (december 1999). NWO-Gebiedsbestuur Geesteswetenschappen.*

digital heritage, humanities and computer science in several ways. The Mission of SURF is to exploit and improve a common advanced ICT infrastructure that will enable higher education institutes better realise their own ambitions and improve the quality of learning, teaching and research. In the SURF Strategic Plan 2003-2006 'The heart of the matter', SURF has changed its perspective radically: the user is now central. With this change, SURF tries to optimise the quality of education and research by applying advanced ICT support where possible. The SURF programme Digital Academic Repositories (DARE) is a joint initiative of the Dutch universities to make all their research results digitally accessible. The KB, the KNAW and NWO are also cooperating in this unique project.

SURF is developing new plans for e-science in the humanities. In a recent report, an attempt has been made to develop a better understanding of those activities and processes in the humanities that are fit for dedicated ICT stimulation and support[19].

5.1.4 MultimediaN

MultimediaN is an initiative of leading researchers in the area of multimedia analysis, database technology, and human computer interaction to improve the scientific base in the Netherlands for applications and services relying on analysis and enrichment of multimedia data. MultimediaN commits itself to a co-ordinated research program based on its current position in the leading edge in multimedia content extraction, efficient multimedia content management, personalised multimedia, and man-machine interaction. The consortium aims to expand and exploit the knowledge in multimedia information systems, standards, interaction, information extraction and condensation, and also in video compression, cognitive assessment of information content, and intelligent interfacing. Results are suited for implementation in the multimedia value chain in its full breadth from content enabling to service delivery.

MultimediaN is conceived as a joint venture with a co-ordinated research program. The form is a virtual centre for knowledge transfer based on multimedia science, where techniques will be demonstrated in prototypes, half-products and first time applications. MultimediaN derives its scientific goals from close interaction with both large national digital archives as emerging high-end multimedia services over (mobile) internet. Every year the CATCH and MultimediaN Programme Committees will jointly organize a seminar, the Dutch Multimedia Event.

5.2 International Context

The CATCH consortium is well aware of the international context. For example: *Het Geheugen* is related to the American Memory project of the Library of Congress, but is more complex, since it does not deal with the collection of the National Library only, but with collections of over 40 museums, archives and libraries. The CATCH-project will of course build on the knowledge from existing international projects. CATCH differs from the Dspace project in that it deals with the massive digital-legacy collections in a wide range of Dutch cultural heritage institutions, while Dspace deals with newly generated digital material only.

[19] *E-based Humanities and E-humanities on a SURF Platform*, Joost Kircz, Kircz Research Amsterdam (2004).

The MIT Media Lab has been very influential in the past in demonstrating on a small scale what is intended to be implemented in a more modern and advanced way, on a very large scale within the CATCH project. Many of our consortium members have close ties with or participate in international projects. Below we deal with several of the projects. We have subdivided the overview as follows: European Union (in 5.2.1), International Networks (in 5.2.2), Related Programmes in the European Union (in 5.2.3), Related Programmes in the World (in 5.2.4).

5.2.1 European Union

'Digital Heritage and Cultural Content' (DigiCULT) is a domain of research activity in the Information Society Technologies (IST) Programme, a European Commission programme addressing the pervasion of Information and Communication Technologies (ICT) into all aspects of the European citizen's life. This programme was already part of the Fifth Framework Programme for Research and Technological Development (RTD) which ran from 1998-2002, and continues to exist as a key thematic priority area within the 6th Framework Programme (2002-2006).

The Work Programme 2003-2004 "Integrating and strengthening the European Research Area in the Community sixth Framework Programme" specifies the content of the activities. "The focus is on improving accessibility, visibility and recognition of the commercial value of Europe's cultural and scientific resources, by developing: advanced digital libraries services, providing high-bandwidth access to distributed and highly interactive repositories of European culture, history and science; environments for intelligent heritage and tourism, re-creating and visualising cultural and scientific objects and sites for enhancing user experience in cultural tourism; advanced tools, platforms and services in support of highly automated digitisation processes and workflows, digital restoration and preservation of film and video material, and digital memory management and exploitation".

With a research focus on eCulture and eScience (i.e., culture and science in a networked environment), DigiCULT aims at establishing a lasting infrastructure of technologies, guidelines, standards, human and institutional networks that will support and extend the role of Europe's libraries, museums and archives in the digital age.

Objectives of the research activities are:

- Enhancing access to and preservation of cultural and scientific heritage resources -particularly those in digital form- thus supporting Europe's heritage institutions and organisations in their core functions,
- Accelerating the appropriation of advanced technologies by Europe's libraries, museums and archives,
- Encouraging convergence in technical approaches and applications for various cultural institutions and networked services by promoting agreement on standards and guidelines critical to managing, preserving and delivering digital cultural and scientific content,
- Fostering increased co-operation between cultural and scientific content holders, i.e. libraries, archives, museums, and the research community or technological application developers, i.e. research centres, academic institutions, ICT companies, etc.

5.2.2 International Networks

In the field of digital cultural heritage, a number of international networks exist, with which the CATCH program will interact and be in contact. Below we mention two of them.

The DELOS Network of Excellence on Digital Libraries[20] **-** Digital Libraries (DL) have been made possible through the integration and use of a number of IC technologies, the availability of digital content on a global scale and a strong demand for users who are now online. They are destined to become essential part of the information infrastructure in the 21st century.

The DELOS network conducts a joint program of activities aimed at integrating and coordinating the ongoing research activities of the major European teams working in DL-related areas with the goal of developing the next generation DL technologies. The objective is to:

- define unifying and comprehensive theories and frameworks over the life-cycle of DL information,
- build interoperable multimodal/multilingual services and integrated content management ranging from the personal to the global for the specialist and the general population. The Network aims at developing generic DL technology to be incorporated into industrial-strength DL Management Systems (DLMSs), offering advanced functionality through reliable and extensible services.

The Network will also disseminate knowledge of DL technologies to many diverse application domains. To this end a Virtual DL Competence Centre has been established which provides specific user communities with access to advanced DL technologies, services, testbeds, and the necessary expertise and knowledge to facilitate their take-up.

The Digital Library Federation (DLF) is a consortium of libraries and related agencies that are pioneering in the use of electronic-information technologies to extend their collections and services. Through its members, the DLF provides leadership for libraries broadly by -

- identifying standards and "best practices" for digital collections and network access,
- coordinating leading-edge research-and-development in libraries' use of electronic-information technology,
- helping start projects and services that libraries need but cannot develop individually.

The DLF operates under the administration umbrella of the Council on Library and Information Resources (CLIR).

5.2.3 Related programmes in the European Union

In the framework of the European Union there are many projects in the cultural-heritage sector. They are certainly interesting but no project coincides with our approach. Below we mention some of the important projects but we refrain from pointing out the differences with the CATCH programme.

[20] http://www.delos.info/

Interoperability
In the 5th Framework, relevant activities were coordinated by the European Commission's Cultural Heritage Applications unit, DG XIII-E2 in Luxembourg. Some activities are HyperMuseum (http://www.hypermuseum.com/), CHIOS (http://www.dl-forum.de/Foerderung/Projekte/CHIOS/), CIDOC (http://www.cidoc.icom.org), META-e (Metadata Engine), SCHEMAS: Forum for Metadata Schema implementers.

Also in the 6th Framework (2002-2006), the European Commission is committed to supporting this area. The research domain "Digital Heritage and Cultural Content" (a research activity in the Information Society Technologies (IST) Programme) will continue to exist as a key thematic priority area within the 6th Framework Programme.

In the domain of semantic interoperability the four most recent programmes in the 5th Framework are CHIMER, COINE, ECHO, and INTERA. Below we provide a brief description.

CHIMER (Children's Heritage Interactive Models for Evolving Repositories; http://dbs.cordis.lu/fep-cgi/srchidadb). CHIMER aims to establish an open international network of children, teachers and museologists for developing an Open Evolving Multimedia Multilingual Digital Heritage Archive as a long-term storage medium for European cultural repositories.

COINE (Cultural Objects in Networked Environments) (http://dbs.cordis.lu/fep-cgi/srchidadb). Empowering European citizens to tell their own stories lies at the heart of the COINE (Cultural Objects in Networked Environments) Project. It will provide the tools needed to create structured, World Wide Web-based environments which are hospitable to local cultural activity but which allow content to be shared locally, regionally, nationally and internationally.

ECHO (European Cultural Heritage Online) (http://echo.mpiwg-berlin.mpg.de, http://www.mpi.nl/echo) is a new project that has as task to provide a rich interdisciplinary access to objects of cultural heritage. Aspects of interoperability at the metadata level between the 4 included disciplines is one of the core aspects.

INTERA: Integrated European language Resource Area is an attempt to solve interoperability problems on a vertical line by creating not only a large metadata domain of language resources, but also by integrating the domain of resource descriptions with those of tool descriptions. The goal is that dependent on the type of selected resources appropriate tools will be selected automatically.

Besides these four programmes, it is relevant to mention TEL.

TEL: The European Library. The objective of TEL is to set up a cooperative framework which will lead to a system for access to the major national and deposit collections (mainly digital, but not precluding paper) in European national libraries. TEL will investigate how to make a mixture of traditional and electronic formats available in a coherent manner to both local and remote users. TEL will contribute to the cultural and scientific knowledge infrastructure within Europe by developing co-operative and concerted approaches to technical and business issues associated with distributed access to large-scale content. It will lay down the policy and develop the technical groundwork for a sustainable pan-European digital library based on distributed digital collections and on the operational digital library developments in the participating libraries and agencies. Project website: http://www.europeanlibrary.org http://www.kb.nl/kb/sbo/netwerk/tel-en.html

For an overview of the many activities in Europe we provide the following list.

CHIOS (Cultural Heritage Interchange Ontology Standardization),
CHLT (Cultural Heritage Language Technologies),
CHOSA (Application of new technology to increase access to the cultural heritage of St. Albans),
CLEF (Cross-Language Evaluation Forum).
COVAX (Contemporary Culture Virtual Archive in XML),
CULTIVATE EU (Cultural Heritage Applications network),
CYCLADES (An open Collaborative Virtual Archive Environment),
DELOS (A Network of Excellence on Digital Libraries),
DOMINICO (On the trace of DOMINICO dell'Allio),
LEAF (Linking and Exploring Authority Files),
MATAHARI (Mobile Access To Artefacts and Heritage At Remote Installations)
MIND (Multimedia International Digital Libraries),
PAST (exPeriencing Archaeology across Space and Time),
POUCE (Portails Culturels Collectifs),
PULMAN (Public Libraries Mobilising Advanced Networks),
PULMAN XT (Extending the European Research Network for Public Libraries, Museums, Archives),
RENARDUS (Academic Subject Gateway Service Europe), and
SANDALYA (An open platform for accessing, co-operatively authoring and publishing the digital heritage of manuscripts and rare books).

Knowledge Enrichment

At the level of manuscripts, an internationally well-known example of cultural-heritage knowledge disclosure is the Electronic Beowulf project. Handwritten manuscripts are presented on-line and are annotated in great detail, disclosing the temporal evolution of the famous Beowulf texts (see further in 5.2.4). This example, however, is one of the few that we consider as exemplary. Many other approaches simply do not address the power of information technology. An example of the latter kind concerns the Historical Archives of the European Communities (http://wwwarc.iue.it/), basically a directory service to physical documents which are only accessible by visiting the archive in persona. A considerably better example is the "Digitale Bibliothek" by the Bayerische Staatsbibliothek, showing transcriptions as well as facsimile images of important printed works (http://mdz.bib-bvb.de). However, navigation is difficult, and no use of hyperlinks from within the images is possible. No panning and zooming facilities are available and the facsimiles are in monochrome black and white. Many projects actually do much worse, merely presenting the facsimiles in a coarse resolution, giving superficial impressions only. A number of 'modern'

European projects do exist, such as MUMIS[21] (Multimedia Indexing and Searching Environment) with an emphasis on streaming media (video).

The COLLATE Collaboratory project[22] comes close to what is ultimately needed in cultural-heritage knowledge disclosure: it "aims at the development and practical usage of a content-centric, user-driven information system for the management of surrogates of fragile historic multimedia objects. As a distributed Web-based multimedia repository, it will function as a 'collaboratory' supporting distributed user groups by dedicated knowledge management facilities such as content-based access, comparison and in-depth indexing/annotation of digitised sources." However, the application examples concern the domain of the cultural heritage of European movies in the 1920s and 1930s. In the audio domain, current technology for content-based retrieval and indexing is quickly developing to a usable level (Zhang & Kuo, 2001)[23]. The European CIMWOS project[24] "aims to facilitate common procedures of archiving and retrieval of audio-visual material. The objective of the project is to develop and integrate a robust unrestricted keyword spotting algorithm and an efficient image spotting algorithm specially designed for digital audio-visual content, leading to the implementation and demonstration of a practical system for efficient retrieval in multimedia databases". This project thus aims at the development of retrieval engines only, without solving the problems of knowledge disclosure around specific high-value objects of the cultural-heritage domain.

In conclusion: although a number of efforts do exist at the European level, the potential for a successful European successor to the Electronic Beowulf approach is much greater if a focused collection from within the Netherlands is used, by researchers from the humanities and from computer science who share a common culture and enthusiasm to preserve it digitally.

Personalisation

There are initiatives on personalisation in the European Union. We provide a few references below. For an example project we refer to the Hermitage Museum's New Web Site.
HyperMuseum (http://www.hypermuseum.com/)
CHIOS (http://www.dl-forum.de/Foerderung/Projekte/CHIOS/)
CIDOC (http://www.cidoc.icom.org)
The Open Heritage initiative (http://www.openheritage.com/intro.html)

5.2.4 Related programmes in the World

There are many international initiatives, most of them of recent date. None of the programmes encountered so far, covers the three themes of the CATCH Programme.

A project to mention is the Hermitage Museum's New Web Site, a cooperation between IBM (Yorktown Heights, NY) and the Hermitage Museum. The project followed the then (1997) visionary ideas of Mikhael Piotrovski, director of the Hermitage. Three end-user applications

[21] *http://parlevink.cs.utwente.nl/projects/mumis/*
[22] *http://www.collate.de/*
[23] *Zhang, T. & Kuo, C.-C.J. (2001). Audio content analysis for on-line audiovisual data segmentation and classification. IEEE Transactions on Speech and Audio Processing, 9(4), pp. 441-457.*
[24] *http://www.xanthi.ilsp.gr/cimwos/*

were identified: (1) multimedia-based art education housed in an education and technology centre, (2) visitor information links, and (3) a new Web site ("that would permit the Hermitage's collections to be searched and better experienced from afar")[25]. For more relevant information worldwide we refer to Kumar et al.[26] In the USA attention is given to the adequate accessibility of The Library of Congress (www.loc.gov).

Another famous and successful pioneering project is "*Electronic Beowulf"* (Kiernan, 1995) on the famous Beowulf manuscripts. In this project, the original handwriting has been scanned in high resolution and has been augmented with a very detailed annotation at both the level of script (the written shapes) and at the level of the textual content. Due to the high quality if this work, the on-line results on Internet and CDROM represent a true form of knowledge disclosure towards experts and regular interested users. A project with a wider scope is represented by "*Digital Scriptorium"* (Faulhaber, 1999). In this latter project, a wide range of mediaeval text is disclosed in digital form, to experts and the general public. The goal of Digital Scriptorium is the knowledge transfer in the area of palaeography (http://sunsite.berkeley.edu/scriptorium/). Fortunately, for the multi-level coding of (a) semantic content, (b) geometric layout structure and (c) typography new standards are emerging, such as TEI (Text Encoding Initiative, http://www.tei-c.org/). These successful international projects may serve as an example for initiatives which are aimed at the preservation of the Dutch cultural heritage.

Finally, we mention the Open Archive Initiative (www.openarchives.org).

[25] *F. Mintzer, G.W. Braudaway, F.P. Giordano, J.C. Lee, K.A. Magerlein, S. D'Auria, A. Ribah, G. Shapir, F. Schiattarella, J. Tolva, and A. Zelenkov (2001). Populating the Hermitage's Museum New Web Site.* Communicaitons of the ACM*, Vol. 44, No. 8, pp. 52-60.*

[26] *Kumar, K.G., et al. The Hot Media architecture: Progressive & Interactive rich media for the Internet. See www.developer.ibm.com/library/articles/hotmedia.html*

APPENDIX I: SIX CORE PROJECTS

Core Project 640.001.401

1a) Project title:
SemanTic Interoperability To access Cultural Heritage

1b) Project acronym
STITCH

1c) Principal investigators
Prof. dr. F. Van Harmelen (Vrije Universiteit)
Drs. H. Matthezing (Koninklijke Bibliotheek)
Dr. P. Wittenburg (Max Planck Institute for Psycholinguistics)

1d) Main project location
Koninklijke Bibliotheek

2) Composition of research team
- 1 Ph.D Student
- 1 Postdoc
- 1 Scientific programmer
- Prof. dr. F. van Harmelen (Vrije Universiteit)
- Drs. H. Matthezing (Koninklijke Bibliotheek)
- Drs. M.C. de Niet (Koninklijke Bibliotheek)
- Prof. dr. G. Schreiber (Vrije Universiteit)
- Dr. P. Wittenburg (Max Planck Institute for Psycholinguistics)

3) Description of the proposed research

3a) Problem statement and research objectives
Cultural-heritage collections are typically indexed with metadata derived from a range of different vocabularies, such as AAT, Iconclass and in-house standards. This presents a problem when one wants to use multiple collections in an interoperable way. In general, it is unrealistic to assume unification of vocabularies. Vocabularies have been developed in many sub-domains, each with their own emphasis and scope. Still, there is significant overlap between the vocabularies used for indexing.

The prime research objective of this subproject is to develop theory, methods and tools for allowing *metadata interoperability through semantic links* between the vocabularies. This research challenge is similar to what is called the "ontology mapping" problem in ontology research.

The overall objective can be divided into three research questions:

1. What kind of semantic links can be identified?
2. Which methods and tools can support manual and semi-automatic identification of semantic links between vocabularies?
3. How can such semantic links be employed to enable interoperable access to multiple collections indexed with heterogeneous vocabularies?

3b) Scientific approach and methodology
The project will be application-oriented. The goal will be to develop methods and tools that can be shown to work for relevant use cases. The project will focus on 19th century cultural-heritage objects in different Dutch collections. For this project we assume that *syntactic* interoperability has been achieved through the representation of metadata and the vocabularies in RDF/OWL format [Brickley and Guha, 2004; McGuinness and van Harmelen, 2004]. This allows the project to zoom in on the semantic interoperability problems.

The project will build on research in ontology mapping. Several authors have proposed mapping relations for use in semantic linking [e.g. Niles and Pease, 2003]. These include equality, equivalence, subclass, instance and domains-specific relations. The project will use these as a starting point and evaluate and extend/revise this set of mapping relations. Research of identification of links will first focus on baseline methods for manual specification of links such as developed within the ICES-KIS 2 project "Multimedia Information Analysis" [Hollink, 2003]. This will be supplemented with techniques from ontology learning targeted at finding such links automatically. The state-of-the-art techniques are not full proof [Handschuh and Staab, 2003], so some form of human validation of the links will need to take place. This is not a big hurdle, as semantic links between vocabularies are a one-time thing. Another technique to consider is the generalization of existing annotations to semantic vocabulary links. For example, if according to a particular annotation the artist of a particular painting belongs to a certain art school, we may hypothesize that this link also exists for other works of the same artist.

With respect to the use of semantic links we will identify a number of typical use cases that should be handled by the tools being developed. Some prototypical use cases are:

- *User sees painting of a historic event, such as the events in Brussels in 1830. She wants information about this event and about other art works concerning this event as well as written witness reports.*
- *User wants to find monuments that constitute particular types of defence works, such as those part of the "Hollandse waterlinie". She also wants information about the architects involved and pointers to writings containing background information.*
- *User wants to find for a particular artist the places where the person lived and worked.*
- *User wants additional information that can be found about certain histories figures (e.g. King William I of The Netherlands or Thorbecke) depicted an a painting?*

These use cases typically require the combination of information from different collection databases.[27] The target user audience for these use cases is the interested lay person.

The following collection databases will be considered for application within the project:

- Catalogue of the Koninklijke Bibliotheek
- Monument preservation
- Army museum
- RKD collection
- Bibliopolis
- Rijksmuseum
- "Geheugen van Nederland" (Memory of The Netherlands)

Vocabularies and thesauri that are of potential interest here include:

- RKD Artist (i.e. Dutch version of ULAN)
- Dutch AAT
- Historic thesauri, such as under development at the Koninklijke Bibliotheek
- Iconclass
- GOO ("Gemeenschappelijke Onderwerpen Ontsluiting"), Koninklijke Bibliotheek
- GTAA (Sound and Vision, see CHOICE subproject)

3c) Scientific relevance

Ontology mapping is becoming an increasingly important research topic. It may provide the background knowledge required for accessing distributed information repositories, both within (large) companies and on the Word Wide Web. Until now, much of the research effort has been spent on making syntactic interoperability feasible, i.e. to represent data models and data in a common (exchange) format. With the advent of XML, and RDF/OWL, these syntactic problems are now (at least in theory) solvable, but this potential is still largely unexplored. Given the fact that semantic interoperability has not been studied very much

[27] This is an indicative list with the aim of making clear the kind of questions this project tries to answer. The project may choose to work on other examples for pragmatic reasons.

yet, this project has taken a use-case driven approach. We expect to show that this technology can be employed to answer a new class of queries over different collections.

3d) Related work

Finnish Museums Online [Hyvonen et al., 2003]:
The joint national museum network developed by the University of Helsinki and The Helsinki Institute for Information Technology HIIT has recently been taken into trial use. The system is based on semantic web technology being seemingly the first of its kind in the world. This project is unique in that it includes a semantic data search system connecting the various collections with each other.

3e) Work programme

The research proceeds in four stages of one year each. Below, the annual planned activities are outlined.

Year 1

- Selection of initial set of collections and vocabularies
- Syntactic transformations to XML/RDF/OWL, where required
- Refinement of initial target use cases into full-blown scenarios
- Construction of baseline manual semantic-linking tool
- First semantic-search prototype

Year 2

- Small-scale user experiments with initial prototype
- Revision of the set of semantic-link primitives
- Facilities for semi-automatic elicitation of semantic links, including generalization from existing annotations
- Second semantic-search prototype

Year 3 & 4

Additional development cycles involving a wider scope of collections, vocabularies and/or use-case functionalities.

3f) Deliverables

D1: Theory of mapping relations required for semantic links between heterogeneous vocabularies

D2: Method and tool for manual identification of semantic links

D3: Algorithms for semi-automatic elicitation of semantic links

D4: Semantic-search tool

4) Expected use of instrumentation

No special equipment is expected to be required.

5) Literature

5a) References to cited work

D. Brickley and R. V. Guha. RDF vocabulary description. Recommendation, W3C Consortium, 10 February 2004. See: http://www.w3.org.

S. Handschuh and S. Staab. Annotation of the shallow and the deep web. In S. Handschuh and S. Staab, editors, Annotation for the Semantic Web, volume 96 of Frontiers in Artificial Intelligence and applications, pages 25-45. IOS Press, Amsterdam, 2003.

E. Hyvonen, S. Kettula, V. Raatikka, S. Saarela, and K. Viljanen. Finnish museums on the semantic web. In Proceedings of WWW2003, Budapest, poster papers, 2003.

D. McGuinness and F. van Harmelen (eds.). OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium, 10 February 2004. Latest version: http://www.w3.org/TR/owl-features/.

Alistair Miles and Brian Matthews. Review of RDF thesaurus work. Deliverable 8.2, version 0.1, SWAD-Europe, 2004. URL: http://www.w3c.rl.ac.uk/SWAD/deliverables/8.2.html.

I. Niles and A. Pease. Linking lexicons and ontologies: Mapping Wordnet to the suggested upper merged ontology. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE ’03), Las Vegas, Nevada, June 23-26 2003.

T. Peterson. Introduction to the Art and Architecture Thesaurus. Oxford University Press, 1994. See also: http://www.getty.edu/research/tools/vocabulary/aat/.

ULAN: Union List of Artist Names. The Getty Foundation. http://www.getty.edu/research/tools/vocabulary/ulan/, 2000.

H. van der Waal. ICONCLASS: An inconographic classification system. Technical report, Royal Dutch Academy of Sciences (KNAW), 1985.

5b) Most important publications of the research team

I. Horrocks, P. F. Patel-Schneider and F. van Harmelen, From SHIQ and RDF to OWL: The Making of a Web Ontology Language, Journal of Web Semantics, 1(1), 2003.

L. Hollink, A. Th. Schreiber, J. Wielemaker, and B. J. Wielinga. Semantic annotation of image collections. In S. Handschuh, M. Koivunen, R. Dieng, and S. Staab, editors, Knowledge Capture 2003 - Proceedings Knowledge Markup and Semantic Annotation Workshop, pages 41-48, 2003.

A. Th. Schreiber, I. I. Blok, D. Carlier, W. P. C. van Gent, J. Hokstam, and U. Roos. A mini-experiment in semantic annotation. In I. Horrocks and J. Hendler, editors, The Semantic Web - ISWC 2002, number 2342 in Lecture Notes in Computer Science, pages 404-408, Berlin, 2002. Springer-Verlag. ISSN 0302-9743.

A. Th. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. IEEE Intelligent Systems, 16(3):66-74, May/June 2001.

J. Wielemaker, A. Th. Schreiber, and B. J. Wielinga. Prolog-based infrastructure for rdf: performance and scalability. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, The Semantic Web - Proceedings ISWC'03, Sanibel Island, Florida}, volume 2870 of Lecture Notes in Computer Science, pages 644-658, Berlin/Heidelberg, October 2003. Sringer Verlag. ISSN 0302-9743.

B. J. Wielinga, A. Th. Schreiber, J. Wielemaker, and J. A. C. Sandberg. From thesaurus to ontology. In Y. Gil, M. Musen, and J. Shavlik, editors, Proceedings 1st International Conference on Knowledge Capture, Victoria, Canada, pages 194-201, New York, 21-23 October 2001. ACM Press.

Core Project 640.001.402

1a) Project title:
CHarting the informatiOn landscape employIng ContExt information

1b) Project acronym
CHOICE

1c) Principal investigators
Dr. M.J.A. Veenstra (Telematica Instituut)
Prof. Dr. G. Schreiber (Vrije Universiteit)
Drs. J.F. Oomen (Nederlands Instituut voor Beeld en Geluid)

1d) Main project location
Nederlands Instituut voor Beeld en Geluid

2) Composition of research team
- 1 Ph.D Student
- 1 Postdoc
- 1 Scientific programmer
- Drs. J.F. Oomen (Nederlands Instituut voor Beeld en Geluid)
- Dr. M.J.A. Veenstra (Telematica Instituut)
- Prof. Dr. G. Schreiber (Vrije Universiteit)
- Dr. P. Wittenburg (Max Planck Instituut for Psycholinguistics)
- Drs. A. Kok (Instituut Collectie Nederland)
- Drs. A van Loo (Nederlands Instituut voor Beeld en Geluid)

3) Description of the proposed research

3a) Problem statement and research objectives
The CATCH research programme will develop key technology to ensure continuous access to the cultural riches of the world. The CHOICE project seeks to chart the uncharted information landscape, focusing on semi-automatic semantic annotation and employing context information.

Semantic annotation involves the annotation of archived objects, such as video, images and books with semantic categories from some standardized metadata repository, such as domain thesauri and ontologies. The use of semantic annotation allows one to widen the search facilities in a collection. For example, annotating a photograph with the semantic category "bed" (in the sense of: to sleep in) from the WordNet thesaurus makes it possible to search for "sleeping beds" while not retrieving other "beds" such as "river beds". As most thesauri have a hierarchical broader/narrower structure, it also makes it possible to generalize or specialize a query in semantic terms: e.g. retrieving photographs of "cribs' (a narrower semantic category) when searching for beds in the "sleeping" sense. Hyvonen (2003) describes an example of a working system in the cultural heritage domain that allows semantic search.

The driving use case of this project is the Sound and Vision video archive. The objective is 1) to show how semantic annotation can be supported in the archiving process by exploiting the available context information and 2) to show how these annotations can subsequently be used to improve search facilities. Hollink et al. (2003) show that linking a number of diverging thesauri to an annotation application for images of paintings can improve both the semantic annotation process for human annotators and the search process. In the CHOICE project, the annotation application developed by Hollink et al. will be adjusted for video annotation. The aim is to construct a video annotation system based on a shared annotation structure (in the Sound and Vision case: iMMix), allowing annotators to mark up video with relevant semantic categories from multiple thesauri relevant for the field.

At the moment automatic techniques for video analysis are still of limited value for the derivation of semantic categories (e.g., Hollink et al., 2004). On the other hand, manual semantic annotation is time-consuming. Therefore, this project will focus on speeding up the manual annotation process by applying natural language processing (NLP) techniques to generate candidate semantic categories that appear in the selected thesauri from (textual) context information. Context information provides peripheral insights into an object; how it was perceived, how it was created, how it relates to other objects made during the same era and so on. Having access to these sources enables users to expand their explorations into greater depth. In the audiovisual realm, examples of sources to be somehow linked to objects include: commentary sheets, external reviews, broadcast schedules, viewer ratings and awards. Within CHOICE, possibly relevant statements and setting descriptions from the textual context information will be offered to the human annotator for approval or rejection. Whether a fragment of the context information is (possibly) relevant for semantic annotation is determined by checking whether concepts from relevant thesauri or from the metadata belonging to the video occur in it. Machine learning and statistical methods for natural language processing and information extraction are applied to determine which terms from fragments or sentences will be used in the statements that are offered to the annotator (Hearst (1999), Jackson and Moulinier (2002), Mitchell (1999).
For the development of a semantic-annotation system for video annotation the following research issues need to be tackled:

1. *How should the annotation interface for images, as developed by Hollink et al., be adapted to video annotation?* In this Sound and Vision case this means integrating the iMMix model into the annotation architecture and incorporating facilities for video browsing and searching, and viewing context information.
2. *Which thesauri and/or ontologies can be used as repositories of relevant semantic categories for archive search?* Typical example corpora could be WordNet, a geographical thesaurus such as TGN, and the "Gemeenschappelijke Thesaurus Audiovisuele Archieven" developed by Sound and Vision and the Filmmuseum.
3. *How can these thesauri/ontologies be partially mapped/integrated?* This issue will build upon the work in the CATCH project STITCH project, also carried out within the CATCH framework.
4. *How can we use NLP and learning techniques to derive relevant semantic categories from the text?* There is a link here to the MITCH project of CATCH.
5. *How can these semantic categorization techniques be used to support the search process?* For example, when searching for video fragments about Limburg, one could use TGN to find geographical parts of Limburg (towns, rivers, lakes, mountains) to enhance the search. As another example, when searching for videos about "crime" it should be possible to find fragments about "murder".

Scoping remarks:

- Allowing all visitors and experts to add additional (semantic) annotation is a avid voluntary cataloguers who will find surprising ways to mine and exploit the treasure trove offered. However, conducting extensive research in this topic is expected to be out of scope for this particular project.
- Integration into the Sound and Vision business process is strictly speaking not part of the project. However, the project will consider business-integration issues that have a general flavor, such as the storage of the actual context information objects and the storage of resulting annotations.

3b) Scientific approach and methodology

The proposed research is methodological. It is aimed at exploiting the possibilities of combining semantic categorization techniques with techniques for natural language processing to make possible semi-automatic semantic annotation. The NLP techniques are provided with relevant concepts (e.g. from thesauri, term lists and metadata) to focus the processing. Thus, the research is not aimed at developing new techniques for natural language processing but on applying existing techniques in a goal-oriented way.
The project will build on existing open standards for data and metadata representation, such as XML and RDF/OWL.

3c) Scientific relevance

The CHOICE project will explore a novel combination of existing semantic categorization techniques and NLP techniques in the context of semantic video annotation. These techniques will be useful in all situations were there are textual annotations of multimedia material and also a set of relevant (possibly heterogeneous) thesauri and/or ontologies. This is a common theme in the cultural-heritage setting. Almost all collections have been annotated with text. In some collections there is some degree of formality because characteristics have already been described with standardized metadata repositories such as AAT. But even in those collections the textual parts may contain relevant parts suitable for semantic search. For example, in painting collections the subject of the painting is typically only described with an informal piece of text. The techniques developed in this project could thus help making semantic subject search possible. A possible use case could be: searching for paintings about fruit will retrieve paintings about apples, pears, grapes, etc.

3d) Related work

CHOICE is a project on the intersection of semantic annotation and natural language processing with an emphasis on (semi-automatic) semantic annotation. CHOICE builds on several projects and work groups the project members are and were involved in with respect to the Semantic Web (e.g, W3C SWBPD[28]), semantic annotation (Hollink et al., 2003, Schreiber et al. 2001), video annotation (IMMix[29]), semantics-based presentation (CHIME[30], Topia[31]) and semantic interoperability (Wittenburg et al. 2004a; 2004b).

Semantic annotation is studied in the semantic-web research field. Both manual techniques and automatic techniques are being used. Annotea[32] is a W3C project targeted at baseline semantic annotation. The CREAM toolset (Handschuh and Staab, 2002b) provides a mix of manual and semi-automatic annotation techniques. The Armadillo approach (Ciravegna et al., 2004) is mainly aimed at using automatic (natural-language) techniques for constructing semantic annotations. These efforts are mainly aimed at text documents. There is relatively little work on semantic annotation of multimedia documents. One of the few examples in the PhD work of Troncy (2003), who did a case study with the archives of INA, the French equivalent of Sound and Vision.

A good overview of current research on semantic annotation van be found in the proceedings of recent Semantic Annotation and Knowledge Markup Workshops (Handschuh et al., 2002a, 2003).

Hyvonen et al. (2003) describe work related to CHOICE an STITCH in the cultural heritage domain. The joint Finnish national museum network developed by the University of Helsinki and The Helsinki Institute for Information Technology HIIT has recently been taken into trial use. The system is based on semantic web technology being seemingly the first of its kind in the world. This project is unique in that it includes a semantic data search system connecting the various collections with each other.

3e) Work programme

The research proceeds in four stages of one year each. Below, the annually planned activities are outlined.

Year 1

Selection of a subset of the Sound and Vision archive well-suited for an early prototype, e.g. because of the availability of relevant thesauri**.** Selection of thesauri. Mapping of thesauri. First version of semantic annotation interface based on the iMMix model.

[28] *Semantic Web Best Practices and Deployment Group: http://www.w3.org/2001/sw/BestPractices/*

[29] *IMMix is a new information system by Netherlands Institute for Sound and Vision, in collaboration with Ministry of Economic Affairs and the Dutch public broadcasters.*

[30] *http://www.niwi.knaw.nl/en/oi/nod/onderzoek/OND1287669/toon*

[31] *http://topia.telin.nl and Rutledge et al. (2003)*

[32] *http://www.w3.org/2001/Annotea*

Year 2
Selection of suitable NLP techniques. Integration of NLP techniques into semantic annotation tool resulting in a second version of the annotation tool. Including semantic search facilities.

Year 3
Exploring the use of the developed techniques outside the Sound and Vision collection, e.g. for the ICN video collection of interviews with painters from the INNCCA project[33] and a linguistic corpus containing audio, video as well as text from MPI. Final version of the semi-automatic semantic annotation tool.

Year 4
Writing of documentation and dissertation.

3f) Deliverables

The project aims to deliver the following products of research:

- Three successive version of a semantic annotation tool
- Conference proceedings papers about the application of NLP techniques in a semantic annotation context etc.
- A Ph.D. thesis

4) Expected use of instrumentation

The team needs sufficient computing power besides normal desktop computers to operate. One high-end computer (dual-CPU, high on memory and permanent storage capactities) will act as computing server.

5) Literature

5a) References to cited work

Fabio Ciravegna, Sam Chapman, Alexiei Dingli and Yorick Wilks, Learning to Harvest Information for the Semantic Web, in Proceedings of the 1st European Semantic Web Symposium, Heraklion, Greece, May 10-12, 2004.

S. Handschuh, S. Staab (eds.). Annotation for the Semantic Web. IOS Press, 2002a

S. Handschuh, M. Koivunen, R. Dieng and S. Staab (eds.): *Knowledge Capture 2003 -- Proceedings Knowledge Markup and Semantic Annotation Workshop*, October 2003

S. Handschuh & S. Staab Authoring and annotation of web pages in CREAM. 11th International conference on World Wide Web Honolulu, Hawaii, USA, pp. 462 - 473 , 2002b. ISBN:1-58113-449-5

Hearst, M. Untangling text data mining. In Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999

Hollink, L., G. Schreiber, J. Wielemaker and B. Wielinga. Semantic Annotation of Image Collections. In S. Handschuh, M. Koivunen, R. Dieng and S. Staab (eds.): Knowledge Capture 2003 -- Proceedings Knowledge Markup and Semantic Annotation Workshop, October 2003.

Hollink, L., G. Nguyen, D. Koelma, G. Schreiber, M. Worring. User Strategies In Video Retrieval: a Case Study. International Conference on Image and Video Retrieval CIVR 2004,Dublin, July 2004.

Hyvonen, E., S. Kettula, V. Raatikka, S. Saarela, and K. Viljanen. Finnish museums on the

[33] *INNCCA is a project of a group of eleven international modern art museums and related institutions. INCCA's most important set of objectives, which are closely interlinked, focuses on the building of a website with underlying databases that will facilitate the exchange of professional knowledge and information about modern art. Furthermore, INCCA partners are involved in a collective effort to gather information directly from artists.*

semantic web. In Proceedings of WWW2003, Budapest, poster papers, 2003.

Jackson, P. and I. Moulinier. Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization. Amsterdam: John Benjamins, 2002.

Mitchell, T. Machine Learning. McGraw-Hill, 1999.

Lloyd Rutledge, Martin Alberink, Rogier Brussee, Stanislav Pokraev, William van Dieten, and Mettina Veenstra. *Finding the Story - Broader Applicability of Semantics and Discourse for Hypermedia Generation*. In: Proceedings of the 14th ACM conference on Hypertext and Hypermedia (pages 67-76), August 23-2003, Nottingham, UK

Guus Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. Ontology-based photo annotation. IEEE Intelligent Systems, May/June 2001.

R. Troncy. Integrating Structure and Semantics into Audio-visual Documents. In: D. Fensel, K. Sycara and J. Mylopoulos (eds.) The Semantic Web - Proceedings ISWC'03, Sanibel Island, Florida. Lecture Notes in Computer Science, volume 2870, Berlin/Heidelber, Springer-Verlag, 2003.

P. Wittenburg, D. Broeder, P. Buitelaar: Towards Metadata Interoperability. Proceedings of the ACL 2004 Conference. To appear. 2004a

Peter Wittenburg, Greg Gulrajani, Daan Broeder, Marcus Uneson:Cross-Disciplinary Integration of Metadata Descriptions. Proceedings of the LREC2004 Conference. To appear. 2004b

5b) Most important publications of the research team

Guus Schreiber, Hans Akkermans, Anjo Anjewierden, Robert de Hoog, Nigel Shadbolt, Walter Van de Velde and Bob Wielinga. Knowledge Engineering and Management: The CommonKADS Methodology, MIT Press, ISBN 0262193000. 2000.

Guus Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. Ontology-based photo annotation. IEEE Intelligent Systems, May/June 2001.

Mike Dean, Guus Schreiber (eds.), Sean Bechofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah McGuinness, Peter Patel-Scheider and Lynn Andrea Stein. OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004.

Lloyd Rutledge, Martin Alberink, Rogier Brussee, Stanislav Pokraev, William van Dieten, and Mettina Veenstra. *Finding the Story - Broader Applicability of Semantics and Discourse for Hypermedia Generation*. In: Proceedings of the 14th ACM conference on Hypertext and Hypermedia (pages 67-76), August 23-2003, Nottingham, UK

P. Wittenburg, D. Broeder, P. Buitelaar: Towards Metadata Interoperability. Proceedings of the ACL 2004 Conference. To appear. 2004a

Core Project 640.002.401

1a) Project title
Reading Images in the Cultural Heritage

1b) Project acronym
RICH

1c) Principal investigator
Prof. dr. E. Postma (Maastricht University)

1d) Main project location
ROB

2) Composition of the research team:

- 1 PhD student (AI, machine learning, and image recognition)
- 1 Postdoc (AI, machine learning, and image recognition)
- 1 Scientific Programmer
- Dr. A.G. Lange (ROB)
- Prof.dr. E. Postma (UM)
- Prof.dr. J. van den Herik (UM)
- Ir. N. Bergboer (UM)
- Drs. E. Drenth (ROB)

3) Description of the proposed research

3a) Problem statement and research objectives

The archaeological heritage covers in time 99% of our collective memory. Its material of study usually lends itself especially to studying everyday life. The scarce remains of our past that are available for study consist mainly of fragmentary and dispersed (parts of) objects. Fundamental in the process of identification of archaeological remains is comparison of the finds with similar objects from elsewhere and recombining the existing knowledge on these objects. To be able to explain archaeological phenomena one compares in first instance (images of) objects at hand with the (images of) objects kept elsewhere. When images match, in depth analysis of descriptions follow and eventually will lead to an enriched knowledgebase.

Archaeology as a discipline has lately seen many changes in the way it is practiced. Under the influence of the new European legislation (Treaty of Valetta, Malta 1992) the number of excavations grew fast. The number of active archaeologists has grown accordingly: from less than 100 before "Malta", to more than 1000 now.
Perhaps the privatisation of field research has the most profound impact. Instead of a situation where excavation and desktop research, policy making and Archaeological Heritage Management were integrated into one or a few rather big institutions, we see the development of an archaeology market with, mainly, small excavation units.

Together these mechanisms put the accumulation of knowledge under severe pressure. Many of the smaller firms have no direct access to the knowledge base, be it in the form of specialist knowledge or in the form of literature. What we see is a stand still in data accumulation and a threshold to the access of knowledge, while the need for ready access to state-of-the-art knowledge is growing at high rate at the same time.

The amount of recovered archaeological objects is beyond our imagination. In the archives and storerooms of the archaeological institutions there are billions of sherds, flints, metal objects, etc. The variation in form, texture (fabric) and decoration has been studied in a

scientific manner for over 200 years. From this collection a corpus of knowledge has been build on the distribution in space and time, the evolution of the technology to make things, and the function and role of particular objects in ancient society. The magnitude of this corpus, partly laid down in books, is nearly just as overwhelming as the number of objects themselves. Because archaeology destroys its own primary sources by excavating, old excavation reports, monographs and catalogues, being the only remaining (secondary) sources, are still essential part of the knowledge base. To communicate all this information archaeologists traditionally use the concept of reference collections. Much like the use of type specimens in biology, archaeologists classify the finds in types and series of types. This is a mental process that combines and recombines evidence and theory from the finds at hand and from earlier archaeological research. The result of this process is usually a theory of the site's socio-economic and cultural role and the presentation of the evidence on which this theory has been build. Sometimes this evidence is presented as a catalogue-like addendum. The ordering of the finds is described and the key objects are depicted in line drawings and photographs. Other researchers may refer to this body of knowledge, make amendments to the interpretation and consequently adjust the classification.
This is what is meant by a reference collection: a constantly updated body of knowledge, consisting of type series, that can be subject of study in itself, but also refer to explicit knowledge accessible in books and implicit knowledge accessible by talking to a specialist, available to all who are interested.

Today we are facing four challenges:
1. How can we safeguard the existing knowledge base?
2. How can we guarantee ready access for all?
3. How can we guarantee the incorporation of new knowledge in a sustainable way?
4. How can we enrich the existing and forthcoming knowledge by new techniques?

To these questions the development of an electronic National Reference Collection (NRc), which is under way, as part of an European wide network of portals to reference collections (eRC) will be an answer. Archaeology is in the first instance firmly and profoundly based on visual inspection and recognition of objects. Images will be central in this development.

The field of digital vision has been developing in such a direction, that now it becomes realistic to incorporate these new techniques into the eRC to enhance the quality of archaeological research and archaeological heritage management in a fundamental way. Automatic recognition of form, fabric, and decoration of physical objects and of printed images is the focus of the RICH-project. This instrument will not only benefit archaeological practice and knowledge building but is of equal importance in education and training.

The results of the RICH project are essential contributions in this development that has as ultimate aims
1. increasing the efficacy and efficiency of digital access to archaeological core knowledge
2. reinforcing the infrastructure on archaeological core knowledge
3. improving the quality of material studies in Dutch archaeological heritage management and archaeological research in Europe, including the formulation of new research area's.

Research question
How can artificial intelligence support the automatic visual analysis of archaeological objects?

3b) **Scientific approach and methodology**
The approach followed in the RICH project is empirical. Machine-learning algorithms are trained on large collections of images. After training, the ability to recognize or classify previously unseen images is assessed yielding a measure of generalisation performance. The scientific methodology employed consists of four phases: (1) data collection, (2) data pre-processing, (3) training, and (4) evaluation.

Data collection. For the archaeological domain, digital data is collected incrementally by digitizing stored objects or newly found objects. Digitization may proceed indirectly by scanning photographs of multiple views of the objects or directly by means of a digital camera. During the project, the size of the digital collection grows steadily. The collection of data is restricted to four classes of objects: pottery, glass, flint and coins. We briefly discuss each of these classes.

- **Pottery.** Often, large collections of pottery are unearthed at archaeological sites. The shapes of the (fragments of) objects obey certain geometrical laws. Together with texture, the shape can be related to a certain period, location, and socio-economical or cultural entity. High-quality classification systems for pottery are available and support the archaeologist in assigning the found object to a certain class. However, the subjective nature of examining the shape and texture of objects hampers the reliability of classification. The pottery project aims at supporting the archaeologist in the classification of unearthed objects by means of advanced visual analysis techniques. It will draw attention both from professional archaeologists and from a potentially wide non-professional audience.
- **Glass.** The late medieval glass collection of the ROB is well classified, dated and documented and consists of a limited number of object shapes. These shapes are often depicted on late-medieval paintings. Archaeologists and art historians are interested to find matches between the documented and depicted shapes because they put constraints on the time and location of the glass under consideration. Using artificial-intelligence techniques, documented two-dimensional drawings or pictures of an object are translated into digital representations of corresponding three-dimensional objects. These representations are matched to the contents of digitized late-medieval paintings in the Rijksmuseum.
- **Flint.** The classification of flint artefacts is a human endeavour. Archaeological experts analyze visual characteristics such as shape and texture to assign the artefact to a certain time and location. In the flint project, a system that is trained to recognize two-dimensional views of flint artefacts is developed along the same lines as in the pottery project. The complex three-dimensional shape of flint artefacts may necessitate a user-guided classification that proceeds as follows. An artefact is presented to a digital camera (under standard light conditions). Using feature-extraction techniques the digital image is transformed and classified with a certain reliability. Initially, the reliability is rather low. However, the user can enhance the reliability by manually rotating the flint artefact in front of the camera until an acceptable classification is achieved.
- **Coins.** Coins are among the most imaginative finds and were collected and studied in the Netherlands, even before Archaeology became a scientific discipline in 1818 at Leiden University[34]. In coins only the illustration is significant. Without having to account for variations in form and texture they are a good starting point for computer vision analysis. For learning and comparison, both digitals images and the coins themselves are in large quantities available at the Koninklijk Penningen en Munten Kabinet.
  The advantages and effects of digitally-guided determination of new coins that are offered by amateur archaeologist should not be underestimated. While it will not replace the expert, it will free him/her from trivial tasks and allows concentrating on more scientific activities. It will have a positive social effect when amateurs can learn about their finds without having to pass thresholds. The net effect will be that much more finds will be reported and that our knowledge will grow tremendously. A similar effect has been noted in Great Britain where the Portable Antiquities Scheme[35] is highly successful.

Data pre-processing. The pre-processing of image data is necessary for three reasons. First, variations in lighting conditions should be minimized as much as possible. The best way to achieve standard lighting conditions is to employ standardize lighting during digitization. Second, noise and sampling artefacts have to be removed to avoid mistakes in the recognition process. Third, the image data has to be transformed into a format suitable for a machine-learning algorithm. A commonly-used method is to apply a wavelet transform in

[34] Brongers, J. A. 2002. Een vroeg begin van de moderne archeologie; Leven en werken van Cas Reuvens (1793-1835). ROB, Amersfoort.

[35] *http://www.finds.org.uk/*

combination with dimension-reduction techniques. The transformation results in what is often called a feature space where distances reflect archaeological similarity of the objects represented.

Training. The feature-space data are submitted to various types of machine-learning algorithms. Each of these algorithms has parameters that need to be optimised. The generalisation performances of each optimised algorithm are compared to assess the most suitable algorithm. Based on experiences in other image-recognition domains, our comparisons will include support-vector machines (Schölkopf and Smola, 2001). and boosting methods (see, e.g., Torralba, Murphy and Freeman, 2004).

Evaluation. Having established the best-performing machine-learning algorithm, the quality of the recognition (or classification) has to be evaluated. For the evaluation process, the judgements of archaeological domain experts are of pivotal importance. Erroneous results may arise for two main reasons. The first reason is technical, i.e., the recognition error is due to limitations in the machine-learning algorithm (or data collection/preparation). The second reason is that existing archaeological classification is inconsistent. Upon erroneous results, it is often difficult to decide whether the cause is technical or domain specific. To resolve errors in an effective way an intensive interaction with domain experts is an absolute necessity.

3c) Scientific Relevance

The project is relevant to both computer science and the cultural-heritage sector. Disclosing the archaeological visual cultural heritage is a major challenge for computer science. The visual-recognition performance of state-of-the-art techniques is limited. We mention three main problems for the automatic recognition of objects. The first problem is the segmentation problem, i.e., the separation of foreground (the object) and background. Often the contours of an object are difficult to recognize automatically due to shadows or partial occlusion. The second problem is the view-dependency of shape. Most natural shapes change when seen from a different viewpoint. Generally, the changes in shape increase with the complexity of the object. An automatic-recognition technique has to achieve view-invariant object recognition (see, e.g., Sung and Poggio, 1998). The third problem concerns variations in texture, lightness and colour. The texture or colour of, for instance, a depicted archaeological object may vary considerably depending on the direction and nature of illumination. These three (and other) problems of visual variance make the automatic recognition of objects difficult. In order to obtain access to the visual cultural heritage, the combination of efficient pre-processing techniques and machine-learning techniques (e.g., support vector machines) has proven to be fruitful.

The proposed project focuses on the development of advanced visual analysis by means of large sets of (digitized) drawings, images or objects of well-determined materials such as pottery, flint, natural stone, coins, and so forth). The automatic analysis proceeds from visual features (such as shape, texture and colour, see Palmer, 1999) that are extracted from the image containing an object (Bergboer, Postma and van den Herik, 2003, 2004). Using specialized machine-learning techniques, the features are mapped onto predefined categories (if available) that may be defined as metadata to enhance efficient search in archaeological image databases. Alternatively, the features may be automatically clustered in a way that is judged meaningful by domain experts which facilitates the development of a classification system.

3d) Related work

The scientific research in the domain of image recognition is manifold. Since the focus of the RICH project is on the interaction with the cultural-heritage domain, it takes existing state-of-the art recognition techniques as a starting point for realising an effective recognition and classification system (Bergboer, Postma, van den Herik, 2003; Torralba, Murphy and Freeman, 2004).

Related work is performed in the context of the ToKeN projects EIDETIC, VINDIT, and AUTHENTIC. These projects address the problems of content-based image retrieval, combined content-based text and image retrieval, and the automatic analysis of visual art, respectively. The already gained and to be gained insights and experiences of these projects are expected to provide a considerable thrust to the RICH project.

3e) Work Programme

The research proceeds in four stages of one year each. Below, the annual planned activities are outlined.

Year 1

The gathering and labelling of large collections of digitized archaeological objects in a limited number of classes. The classes include pottery, glass, and flint. The post-doc, Ph.D. researcher, and scientific programmer collaborate with domain experts in determining the best procedure for setting up the digital collection. The post-doc focuses on acquiring relevant domain knowledge and setting up the experimental environment. The Ph.D. researcher surveys the scientific literature on relevant pre-processing (feature-extraction) and learning techniques.

Year 2

The Ph.D. and post-doc researchers experiment with various feature-extraction approaches to build a suitable feature space for the type of object and task at hand.
The scientific programmer combines and rewrites existing software for pre-processing and machine learning. The Ph.D. and post-doc researchers apply supervised and unsupervised learning techniques to achieve automatic labelling (or generation of metadata) and clustering, respectively. Archaeological experts are closely involved in the assessment of the approach and the results obtained.

Year 3

The scientific programmer refines the techniques (and hardware) to obtain a reliable system for indoor and outdoor application. Field tests and empirical validation of the system will be performed by the Ph.D. and post-doc researchers. The possibilities of enhancing the classification performance and reliability through user-guided classification are explored.

Year 4

Delivering a automatic interactive visual-analysis system for the archaeological objects.
Delivering the underlying software tailored to the in different cultural-heritage institutions.
Delivering a final report (Ph.D. thesis, guide to the software, and scientific papers).

Throughout the duration of their appointments, the researchers deliver reports and scientific papers on their (intermediate) results. Domain experts from the ROB and other application domains will be involved at various stages during the project. A preliminary assessment will be made of the suitability of the system for application to the Galapagos and butterfly-wings datasets of Naturalis. All software will be made available to the ROB and other cultural-heritage and academic institutes involved in the CATCH project.

4) Expected use of instrumentation

The RICH project is expected to use high-performance desktop PCs (3GHz, 1GB RAM, > 100 GB HD memory). A DVD reader-writer allows for long-term storage of acquired data. In addition, high-quality capturing hardware (scanner and/or digital camera) and supporting lighting materials are required.

5) Literature

5a) References

N.H. Bergboer, E.O. Postma and H.J. van den Herik (2004). A Context-Based Model of Attention. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)* (accepted for publication).

N.H. Bergboer, E.O. Postma and H.J. van den Herik (2003). Context-Enhanced Object Detection in Natural Images. In *Proceedings of the Belgian-Netherlands AI Conference (BNAIC) 2003*, pages 27-34, October 2003, Nijmegen, The Netherlands

S. E. Palmer (1999). Vision Science, Photons to Phenomenology. MIT Press, Cambridge, Massachusetts.

B. Schölkopf and A. J. Smola (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA: MIT Press.

K. K. Sung and T. Poggio (1998). Example-based learning for view-based human face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1):39–51.

A. Torralba and A. Oliva. (2003). Statistics of Natural Image Categories. *Network: Computation in Neural Systems*, Vol. 14, 391-412.

A. Torralba, K. Murphy and W. Freeman (2004). Sharing features: efficient boosting procedures for multiclass object detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (in press).

5b) Five most important publications of the research team

András, P., Postma, E.O., and Herik, H.J. van den (2001). Natural Dynamics and Neural Networks: Searching for Efficient Preying Dynamics in a Virtual World. Journal of Intelligent Systems, 3(3), 173-202.

Bergboer, N.H., Postma, E.O., and Herik, H.J. van den (2004). A Context-Based Model of Attention. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)* (accepted for publication).

Herik, H.J. van den and Postma, E.O. (2000). Discovering the Visual Signature of Painters. In N. Kasabov (Ed.), *Future Directions for Intelligent Systems and Information Sciences. The Future of Speech and Image Technologies, Brain Computers, WWW, and Bioinformatics* (pp. 129-147). Physica Verlag (Springer-Verlag), Heidelberg-Tokyo-New York.

Kortmann, R., Postma, E.O., and Herik, H.J. van den (2001). Evolution of visual systems constrained by a resolution trade-off. Artificial Life (special issue on the evolution of sensors in nature, hardware and simulation) 7 (2), 125-145.

Postma, E.O., Herik, H.J. van den, & Hudson, P.T.W. (1997). SCAN: a scalable neural model of covert attention. *Neural Networks*, 10 (6), 993-1015.

Core Project 640.002.402

1a) Project title
Script Analysis Tools for the Cultural Heritage

1b) Project acronym
SCRATCH

1c) Principal investigator
prof. dr. L.R.B. Schomaker (Kunstmatige Intelligentie/Rijksuniversiteit Groningen)

1d) Main project location
Nationaal Archief

2) Composition of the research team

- PhD student (handwriting classification)
- Postdoc (layout and language modelling)
- 1 Scientific programmer
- drs. Cathy Jager (Nationaal Archief)
- Jacques Bogaarts (Nationaal Archief)
- prof. dr. Lambert Schomaker (KI/RuG)
- prof. dr. John Nerbonne (CLCG/RuG)
- drs. Katrin Franke (KI/RuG)
- drs. Marius Bulacu (KI/RuG)

3) Description of the proposed research
Large collections of handwritten material do not lend themselves easily to simple access on the basis of keywords or traditional information-retrieval methods. For human readers, it is difficult to read the handwriting of another person, and this is even more difficult if the writing originates from a different period in history. Under these conditions it may be appreciated that the *automatic* recognition of handwriting, i.e., the automatic conversion of a text image to a coded representation in ASCII or Unicode, is a major research problem.

Current technology in optical character recognition (OCR) is primary aimed at handling well-separated characters in a known and neat machine-type font for office applications. Current and historical handwriting styles cannot be recognized with sufficient accuracy, in particular if the data consist of cursive-connected styles. In cursive-connected script, the segmentation into characters is a problem in itself, to the extent that "optical word recognition" (OWR) would have been a more suitable acronym. Current experiences with the automatic recognition of cursive script have shown that the original objective of automatic handwriting recognition, i.e., a strict left-to-right *veridical* transcription of a scanned page of handwriting, cannot be reached within the next decade. However, if the goals are defined in a more realistic way, current technology may play an important role in providing tools for retrieval and semi-automated methods of script annotation.

Recent developments in machine learning allow for self-organized categorization of shape content, which can be used for search in large script-image data bases. Although it will not be possible to automatically obtain a flawless transcription of handwritten material, it may very likely be possible to select a paragraph of handwritten text by means of a computer mouse and ask for similar content in a huge collection. If users are willing to enter relevant keywords during this process, it will be possible to apply state-of-the art machine learning to associate coded (ASCII) text and handwritten shape. In such a system concept, the accuracy of text retrieval will improve as a function of the number of users and the number of queries.

The research question then is, whether the application of search and semi-automated annotation tools will provide an effective improvement over purely manual transcription methods. Furthermore, current methods in handwriting analysis systems may allow for new ways of search and retrieval. Traditionally, there is a focus on the textual content of a manuscript, but advanced techniques may allow for a detailed in-depth analysis of the character shapes as well. On the basis of the proposed research, it will be possible for a user in the near future, to pose queries concerning writer identity, writer age, writer schooling, as well as posing questions concerning the type of writing implement or paper quality. In order to solve the problems of accessing handwritten document collections, methods are needed from several research domains: image processing, pattern recognition for shape classification; layout modeling and content metadata research as well as stochastic modeling methods from computer linguistics. A staged research process will be proposed, starting with clean and homogeneous handwritten collections, while ending with an evaluation of the developed methods on more difficult heterogenous script collections at the end of the project.

Relevance for the National Archive and beyond
The number and volume of handwritten collections is huge. As an estimate, the hand-written script collections make up 99% of all available documents of historical interest. Digitization in the form of scanning is futile: It will produce unwieldy digital image collections of which the textual content can only be read page by page, by a human user. In the cases where digitization actually has taken place, the indexing is either crude and superficial (and therefore of limited use) or based on a detailed manual transcription (and therefore extremely expensive). Research which is directed at the development of algorithms for semi-automatic annotation, search and retrieval of text in handwritten collections will be of an extremely high relevance. The resulting tools will be useful in a wide range of related problems concerning handwritten archives.

Related project at National Archive
Document analysis "*Kabinet van de Koningin"* (Nationaal Archief)
The "Cabinet of the Queen" archive concerns a large collection of documents which are diverse in nature but which are also clearly delineated in terms of topicality, historical period, relevant actors etc. This, mostly handwritten, collection consists of handwritten index books which refer to handwritten summary and reference books, which in turn refer to archived boxes containing documents and handwritten letters by ministers and the queen. The number of archivist writers is limited. These professionals produced the handwritten index structure in a clear and regular script style which lends itself excellently as the basis for the development of SCRipt Analysis Tools for the Cultural Heritage (SCRATCH). The size of the collection is important (several tens of thousands of pages). The historical importance of this archive is considerable, covering late nineteenth and early twentieth century, thus offering a unique view on the details of the Dutch constitutional monarchy at work in a parliamentary democracy. By virtue of the systematic manner in which this collection was constructed, it constitutes an ideal starting point for the design of expert-support system tools for annotation *(quadrant B),* text-content based data mining *(quadrant B)* as well as structural analyses *(quadrant C*). The basic research perspective is that of knowledge enrichment.

3b) Scientific approach and methodology
The information content within handwritten patterns at the local level of characters and syllables is limited. For this reason additional information is needed, as is the case in automatic speech recognition. If additional knowledge is available, it provides useful constraints to narrow down the number of possible text interpretations of a fragment . In its most simple form, context knowledge concerns a list of words to be expected in the domain of a collection (e.g. administrative and political topics). However, in large and open domains, the lexical constraints are insufficient to allow for reliable word classification. Under such conditions (stochastic) language models are usually applied, capturing syntactic and semantic regularities in the input streams. Additionally, in the case of handwriting recognition, an important source of constraints is formed by the two-dimensional space of textual elements and their structural relationships. As an example, the automatic recognition of addresses on postal envelopes only yielded useable text-recognition performances after knowledge models were developed which correlate the text layout and its content. Similarly, in the recognition of handwritten historical documents, it will be necessary to integrate the

following four sources of information: handwritten shapes, linguistic models, layout models and domain knowledge. The "Cabinet of the Queen" archive constitutes an excellent starting point for the development of a system architecture consisting of individual tools which allow for pre-processing and annotation, while exploiting the underlying two-dimensional (2-D) layout grammar as well as knowledge from the content-domain. Rather than aiming at a single solution for this particular collection, the goal is to develop a generic methodology which allows for the modelling, annotation and ultimately, recognition of handwritten materials in any historical collection. Thus, the proposed project consists of several design and evaluation phases:

1. Layout/content analysis, index-structure modelling
2. Text-image pre-processing and segmentation tools
3. Text-feature extraction and machine learning (clustering & Kohonen maps)
4. Tools for (semi)automated annotation and system training
5. Empirical evaluations, applying the Information-Retrieval paradigm
6. Upscaling to large numbers of documents
7. Generalisation to more difficult collections

The availability of methods for using handwritten indices or handwritten terms within thesauri will allow for a broadening and deepening of the research results of the first phase of the project. Possibilities include: (a) content-based clustering which is coupled to metadata clusters, leading to a propagation of content labels from machine text to handwritten sections and vice versa (b) continuous learning through annotation, and (c) cross-collection impact of methods for the analysis of numerical handwritten material, date formats, proper names and geographical names etc. By using "Provincial Archives", the robustness of the proposes approach can be evaluated on these difficult and heterogeneous collections. The most difficult collection, in this respect, is the "Afrika ex Artis" collection, which is heterogeneous in handwritten style, chaotic in layout patterns and is mixed with drawings as well as multiple-colored editing corrections. It is an open research question which modules of the developed tool kit will be applicable on this very difficult collection at the end of the project. It is expected that the yield for the administrative collections will be higher in any case.

3c) Scientific relevance
The design of Reading Systems for historical document collections has become an area of increased activity within the scientific community which is represented by Technical Committee 11 (TC-11) within the International Association for Pattern Recognition. The difficulty of recognizing cursive script (Steinherz et al., 1999) and the specific requirements from within the application domain have spawned cross-fertilization between (a) traditional Handwriting Recognition research, (b) Information Retrieval and (c) Knowledge Engineering. A central recurring theme in handwriting-recognition research is the balance between the amount of structure which can be uncovered algorithmically and the amount of structure which is engineered into a system as formalized human knowledge. It has become apparent that there is an upper limit to the amount of structure which can be detected using statistical methods, due to the intrinsic variation and variability of script. For example, it would not be feasible to apply a grammar-induction engine to one million scanned pages and expect that a perfect shape-language grammar will be the result. However, by realistically starting with a minimum amount of engineered knowledge structure and a minimum amount of supervised training, state-of-the art statistical-learning systems may be bootstrapped to yield interesting results. The recognition of connected-cursive script remains an ultimate challenge to Pattern Recognition and Artificial Intelligence: Word-classification performances are much lower than in automatic speech recognition, where the scientific community is at least ten times as large as in cursive-handwriting recognition. The presence of the human reader remains to constitute a challenge as well: functional Reading Systems do exist!

3d) Related work
Since it has become apparent that a veridical automated left-to-right transcription of handwritten collections is not feasible, researchers have identified Information Retrieval (IR) (Salton et al., 1975) as an alternative approach to computer-based processing of handwritten material. Here, IR concerns an application context where users are regularly

querying a document collection and are willing to label the retrieved patterns. For example, a promising route is the use of word spotting (Lavrenko et al., 2004) on the basis of holistic word patterns. This approach is opposed to character recognition, which would be ill posed in, e.g., sloppy or complex script. New feature schemes (Rath & Manmatha, 2003a) are combined with existing matching methods (Rath & Manmatha, 2003b) to enable "Googling", i.e., keyword-based search in large collections. At the same time, handwritten documents can be clustered in ways which are similar to the "*bag of words*" approach in traditional Saltonian Information Retrieval (Rath et al., 2004; Nicolas et al., 2003). In our case, it would be more appropriate to use the expression "*bag of glyphs*", referring to a shape-frequency vector for a document. Both shape-based and text-based indexing are possible today as the basis for what could be dubbed a SR (Script Retrieval) system. Manmatha & Rath (2003) and Govindaraju & Xue (2004) describe methods for convenient index construction. Particularly advantageous is the condition where a partial ("Unicode") transcription of sufficient size exists to correlate transcribed words with handwritten word shapes (Tomai et al., 2002). Several studies address the problems of knowledge engineering (Feldbach & Tönnies, 2003). Here, detailed layout and content constraints allow for a tremendous increase in system performance, at the cost of initial human expert-knowledge input. Last but not least, an important topic concerns the user goals and requirements, which ultimately determine the usability of the envisage pattern-recognition and retrieval algorithms (Nicolas et al., 2003). The proposed project intends to exploit the available expertise in handwriting recognition and machine learning in the project team and combine this with concepts from information retrieval and layout modelling in order to develop new robust and generic methods for handwriting annotation and retrieval.

3e) Work Programme

Year 1

General framework development, concerning metadata standards, image pre-processing, annotation and indexing structure. Cooperation with the metadata researchers at this stage is conducive. The CH partner acts as an interface in this respect. The PhD students are acquainted with the raw materials and operating procedures at the Nationaal Archief. From the onset, it will be made continuously clear that generic solutions are to be preferred above engineered constructions with a local overfit to the problem.

Year 2

Developing and evaluation of individual methods at the levels of language and content, layout and handwritten text classification. Exemplary machine-learning solutions will be presented, using limited data sets (indices and actual documents). A work-flow definition for the annotation of fresh materials will be proposed. Results from the parallel metadata consortium (STITCH) will be incorporated, where appropriate.

Year 3

Scaling up and diversification. Using the tools that have been developed in the first phase of the project, their behaviour and performance will be assessed as a function of the amounts of image material and the diversity in terms of layout and script styles. If the proposed working methods are sufficiently generic, there will exist a cookbook for the processing pipeline from scanning, pre-processing, annotation to classification training, this time applied to an unseen collection of larger heterogeneity than the original "Cabinet of the Queen" archive. Where possible, collaboration with the MITCH project on text mining will be realized.

Year 4

Rounding up, writing of documentation and dissertation. Dissemination of results within the Nationaal Archief and other interested partners within the Cultural Heritage sector.

4) **Expected use of instrumentation**

The methods developed in this project will run on desktop PCs with 1GB internal memory, 3 GHz minimum speed and 120 GB minimum hard-disk capacity. High-quality DVD (RW) long-term storage will be needed. Part of the budget will concern the transfer and conversion of text images from the cultural-heritage to the research groups.

5) Literature

5a) References

M. Feldbach & K.D. Tönnies (2003). *Word Segmentation of Handwritten Dates in Historical Documents by Combining Semantic A-Priori-Knowledge with Local Features*, Proceedings of the 7th Int. Conference on Document Analysis and Recognition}, ISBN 0-7695-1960-1, 333-337, IEEE Computer Society.

V. Govindaraju & H. Xue (2004). Fast Handwriting Recognition for Indexing Historical Documents. Proc. of DIAL'04, pp. 314-320.

V. Lavrenko, T. M. Rath & R. Manmatha (2004). Holistic Word Recognition for Handwritten Historical Document. Proc. of the Int'l Workshop on Document Image Analysis for Libraries (DIAL), Palo Alto, CA, January 23-24, pp. 278-287.

R. Manmatha & T. M. Rath (2003). Indexing of Handwritten Historical Documents - Recent Progress. Proc. of the 2003 Symposium on Document Image Understanding Technology (SDIUT), Greenbelt, MD, April 9-11, pp. 77-85.

S. Nicolas, Th. Paquet, L. Heutte (2003). Digitizing cultural heritage manuscripts: the Bovary project. ACM Symposium on Document Engineering 2003, pp. 55-57

T. M. Rath, R. Manmatha & V. Lavrenko (2004). A Search Engine for Historical Manuscript Images. Proc. of the ACM SIGIR 2004 Conf., Sheffield, UK, July 25-29. [in press]

T. M. Rath & R. Manmatha (2003a). Features for Word Spotting in Historical Manuscripts. Proc. of the 7th Int'l Conf. on Document Analysis and Recognition (ICDAR), Edinburgh, Scotland, August 3-6, Vol. 1, pp. 218-222.

T. M. Rath & R. Manmatha (2003b). Word Image Matching Using Dynamic Time Warping. Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, June 18-20, Vol. 2, pp. 521-527.

G. Salton, A. Wang, and C. Yang (1975). A Vector Space Model for Information Retrieval. Journal of the American Society for Information Science, volume 18, pp. 613-620.

T. Steinherz, E. Rivlin, E., N. Intrator, N.(1999). Offline Cursive Script Word Recognition: A Survey, IJDAR, Vol. 2(3), pp. 90-110.

C.I. Tomai, B. Zhang & V. Govindaraju (2002). Transcript Mapping for Historic Handwritten Document Images. Proc. of the Eighth International Workshop on frontiers in Handwriting Recognition (IWFHR-8), Niagara-on-the-Lake, Ontario, Canada, August 6-8.

5b) Five most important publications of the research team

L. Schomaker & M. Bulacu (2004). Automatic writer identification using connected-component contours and edge-based features of upper-case Western script. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 26(6)*, pp. 787 - 798.

Schomaker, L.R.B. (1993). Using Stroke- or Character-based Self-organizing Maps in the Recognition of On-line, Connected Cursive Script. Pattern Recognition, 26(3), 443-450.

Louis Vuurpijl, Lambert Schomaker & Merijn van Erp (2003). Architectures for detecting and solving conflicts: two-stage classification and support vector classifiers Int. Journal on Document Analysis and Recognition (IJDAR) Vol. 5(4), pp. 213 – 223.

Schomaker, L.R.B. (1998). From handwriting analysis to pen-computer applications. IEE Electronics Communication Engineering Journal, 10(3), pp. 93-102.

Helsper, E., Schomaker, L., & Teulings, H.L. (1993). Tools for the recognition of handwritten historical documents. History and Computing, 5(2), 88-93.

Core Project 640.002.403

1a) Project title

Mining for Information in Texts from the Cultural Heritage

1b) Project acronym

MITCH

1c) Principal investigator

Dr. A.P.J. van den Bosch (Tilburg University)

1d) Main project location

Naturalis

2) Composition of the research team

- 1 Ph.D. student
- 1 postdoc
- 1 scientific programmer
- D. Houtgraaf (Naturalis)
- J. van Tol (Naturalis)
- Dr. A.P.J. van den Bosch (Tilburg University)
- Prof. dr. W.M.P. Daelemans (Tilburg University and University of Antwerp)

3) Description of the proposed research

3a) Problem statement and research objectives

Text mining, a research domain of natural language engineering (an interdisciplinary field of computer science and linguistics), has advanced to a level at which automatic language technology and information extraction modules can be applied to vast amounts of text and analyse these texts on syntax, document structure, and topical-semantic information such as named entities, propositions, relations, and topics (Hearst, 1999; Jackson and Moulinier, 2002). These methods, based on statistical models and machine learning models from artificial intelligence (Mitchell, 1997), are robust and fast. Training material as well as trained systems are currently available for the analysis of Dutch texts (Van den Bosch and Daelemans, 1999; Van Halteren et al, 2001; Hendrickx and Van den Bosch, 2001, 2003). New systems could be trained and tuned to a particular domain of Dutch text easily. The more restricted the language use in a domain is, the better and easier the learning of such domain-specific modules becomes.

There is no intrinsic bound to the type of text that could be analysed by these methods. Text can be "ungrammatical" or even be a list of terms stored in database records. The more structure a collection of texts has, the more possibilities there are for machine learning systems to learn the regularities or syntax of the structure, and apply it to new data or find inconsistencies in existing structured data.

There are also no intrinsic restrictions on the morphological, syntactic, and semantic structures that could be learned; they could be general (for example, find all proper names in a document and determine whether they are persons, organisations, or locations, Tjong Kim Sang, 2002), or they could be tailored to a domain in which particular types of entities and facts should be found and labeled. Once these generic or specific methods are developed, they can be used for supporting further annotation, fully automatic annotation, or the automatic discovery of inconsistencies in previously labeled material.

The research question thus reads: how can language technology and text technology support the automisation of knowledge enrichment and understanding of digitised cultural-heritage texts and textual object data bases? How scalable and robust are these techniques in analysing sentence and text structure syntactically and semantically-informationally?

3b) Scientific approach and methodology

The proposed research is empirical – in essence it measures the generalization performance of learning algorithms applied to targeted problems, to obtain estimates on the generalization performance of the system when applied to amounts of unseen cases of the same problem automatically. A problem in textual analysis or in data base consistency checking can always be formulated as a learnable classification task (Daelemans, 1995), learnable by any classifier-learner from machine learning (Mitchell, 1997). The basic element of the machine-learning approach is the *experiment*, in which one particular learning algorithm is trained on a set of classified training examples, producing a learned model of the classification task. This model is then applied to a held-out set of test examples, producing classifications for all test examples. Given a reference labelling for these examples, standard evaluations (accuracy, precision, recall, F-score) and any other derived domain-specific evaluation metric) can be applied to these experimental outcomes, providing information on the success of the method (at least compared to well-chosen baseline performance scores). Comparisons among outcomes of cross-validation experiments (e.g. of different machine learning algorithms applied to the same training and test data), in which training and test material are systematically varied, furthermore allow for statistical significance tests.

Scaling aspects of the application of machine learning techniques to very large datasets are investigated through *learning curve* experiments (cf. Van den Bosch and Buchholz, 2002). Learning curves, i.e. systematic measurements of generalization performance when increasing amounts of learning material are available, provide indications whether a learning plateau is reached (and no more learning material needs to be labelled for training), or more labeled data would provide increased performance.

3c) Scientific relevance

The proposed research builds on existing work, but extends it in two innovative ways. First, although natural language processing systems such as parsers have been available for decades, there have been and continue to be robustness problems in some real-world applications; the usage of statistical techniques and machine learning has changed this dramatically during the past decade, and the Tilburg research group has been instrumental in this international change.

Second, the proposal aims at semantics (meaning, informational structure, and intention) as much as at syntactic structure, where the latter has been the focus of most earlier work. Semantics, certainly in the cultural-heritage world, is more central to human interest as meta data than syntactic structure, which not only changes through time, but also merely is the carrying medium for information and meaning. The current state of the art is that worldwide a range of annotation projects is performed on different types of textual data to enable a breakthrough in the development of semantic processing in the next decade. The current proposal aims to contribute directly to this international development.

3d) Related work

The proposed project is tightly integrated with other current projects of the ILK[36] (Induction of Linguistic Knowledge) research group of Tilburg University (led by Van den Bosch and Daelemans), and it builds on the well-established basis of this group in terms of expertise, research methodology, and software for machine learning of natural language processing and data mining. The ILK group is taking part in IMIX[37] (Interactive Multimodal Information Extraction), an NWO programme focusing on the development of domain-specific question-answering. ILK is supplying IMIX with a robust semantic tagging module that, in a closed domain, can analyse texts on concepts, relations and topics. In another related strand of work, ILK has been closely involved in automatic and expert-supporting corpus annotation systems for the 10-million-word CGN[38] (Spoken Dutch Corpus) (Oostdijk and Broeder,

[36] *http://ilk.uvt.nl*

[37] *http://www.nwo.nl/imix*

[38] *http://lands.let.kun.nl/cgn/ehome.htm*

2003). A third related strand of work is hybrid image-text retrieval in the TOKEN-2000 project VindIT.

ILK actively maintains TiMBL, a home-grown machine learning package with a wide international[39] user group, along with customized TiMBL-based natural language processing and text analysis software for part-of-speech tagging (Van Halteren et al., 2002), morphological analysis (Van den Bosch and Daelemans, 1999), parsing (Daelemans et al., 1999), word sense disambiguation (Hoste et al., 2002), named-entity recognition in various languages including Dutch (Buchholz and Van den Bosch, 2002; Hendrickx and Van den Bosch, 2003), and information extraction (Zavrel and Daelemans, 2003).

Related work outside the Tilburg research group with a relation to the group includes two European projects at the CNTS[40] (Center for Dutch Language and Speech) at the University of Antwerp, led by Daelemans: BioMinT, on biological text mining, and MUSA, on automatic subtitling, both using robust language technology developed in cooperation with the Tilburg group. The Antwerp group is also a member of the new European Network of Excellence PASCAL[41] (Pattern Analysis, Statistical Modelling and Computational Learning), which hosts a relevant "Information Retrieval & Textual Information Access" programme.

Although there is ample work on customized text mining in restricted commercially relevant domains such as bioinformatics and pharmaceutics, related work in the area of text mining for cultural heritage is sparse. Likely the closest related project to the one proposed here is VIADOCS[42], a joint effort of the Natural History Museum (London, UK) and the University of Essex. VIADOCS is aimed at computerizing archive card indexes (either hand-written or typed) to support data-based research initiatives in the biodiversity research area. Much like the project proposed here, VIADOCS involves a close cooperation among the parties. Naturaiis has contacts with the researchers involved in this project, and more contact will be sought.

A second related project is the Perseus Digital Library[43] (main site Tufts University), which performs automatic indexing on a vast collection of digitized electronic documents by named entity recognition and date parsing. Perseus is a member of the umbrella network CHTL[44] (Cultural Heritage Language Technologies), specifically aimed at Greek, Latin and Old Norse texts. The network maintains "work packages" for morphological analysis, document clustering, and term extraction. Its presence on the web is a good example of how the currently proposed project could be publicized.

Technology-wise, the European project DOT.KOM[45] and the European Network of Excellence PASCAL[46] host research in text mining and information extraction. They are exemplary of the multidisciplinarity of the field, combining research teams from information retrieval, natural language processing, and knowledge management – much like the present proposal.

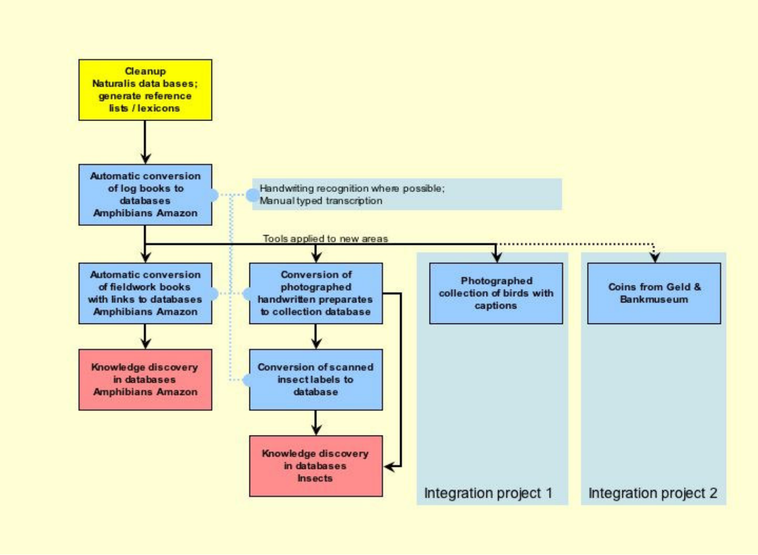
3e) Work programme

The three-person team will follow a primary research line that centres around Naturalis data, in particular logbooks and fieldwork books on amphibians from the Amazon region. This data is partly digitized, but largely handwritten, thus in need of manual typed transcription or automatic handwriting recognition (for which we will seek cooperation with the SCRATCH project). The second dataset for the primary research line is insect data, both in the form of handwritten notes on microscope sections, and small labels attached to insect specimens. In both primary-line knowledge enrichment projects, the end goal is to arrive at automatic

[40] *http://cnts.uia.ac.be/*
[41] *http://www.pascal-network.org*
[42] *http://www.essex.ac.uk/ese/research/vasa/viadocs/*
[43] *http://www.perseus.tufts.edu/*
[44] *http://www.chlt.org/*
[45] *http://kmi.open.ac.uk/projects/dotkom/*
[46] *http://www.pascal-network.org*

knowledge discovery in both areas. This knowledge discovery should be facilitated by the following basic steps:

1. Expert-assisting and automatic cleanup of data (logbook, labels, fieldwork books), including data base fields of particular interest (dates, locations, authors)
2. Linking of terms and phrases to reference books (on amphibians and insects), ontologies, and nomenclature (with a possible link to ongoing work in the STITCH project)
3. Linking of terms, phrases, and domain-specific data base fields (dates, locations) to background texts (encyclopaedia, travel books)


Cleanup Naturalis data bases; generate reference lists / lexicons
Automatic conversion of log books to databases Amphibians Amazon
Handwriting recognition where possible; Manual typed transcription
Tools applied to new areas
Automatic conversion of fieldwork books with links to databases Amphibians Amazon
Conversion of photographed handwritten preparates to collection database
Photographed collection of birds with captions
Coins from Geld & Bankmuseum
Knowledge discovery in databases Amphibians Amazon
Conversion of scanned insect labels to database
Knowledge discovery in databases Insects
Integration project 1
Integration project 2

The planning over time is as follows.

Year 1

The team collects an inventory of standard language technology methods (from machine learning: memory-based learning, support vector machines, maximum entropy models, markovian models, rule induction methods, and hybrids) for (1) tokenization, spelling checking, and normalization; (2) morphological analysis and parsing of Dutch text; (3) named entity (person names, locations, dates) recognition in Dutch text. Methods are used to cleanup data semi-automatically (in GUI tools assisting experts) and fully automatically. Analysis of typed-transcribed "Amphibians in Amazon" logbooks. Automatic conversion of log books to databases. Results are disseminated through reports and conference or journal papers.

- Ph.D. student: apply LT methods to incoming transcribed data; conversion to database
- Postdoc: LT technology toolbox collection; set up cleanup project; assist Ph.D. student in setup; project supervision
- Scientific programmer: create software environment; create/adapt annotation GUI with database integration

Year 2

Area 2 (Insects) is included in the study. Cleanup and database conversion of area 2 is initiated, while work on problem area 1 continues. Modules are made operable for integration. Experts are confronted with developed methods, and are asked to run tests and evaluations. Results are disseminated through the usual channels.

- Ph.D. student: apply LT methods to incoming transcribed data; conversion to database
- Postdoc: assist Ph.D. student in experiments; project supervision
- Scientific programmer: make modules integratable with other CATCH project software

Year 3

Continued work on developing the core demonstrator, that showcases the two areas (amphibians and insects). If possible, existing modules are applied to data in new problem as data from other teams (semantic annotation in STITCH, script analysis in SCRATCH, and image processing in RICH) comes in. At the end of year 3 the Naturalis software showcase demonstrator is released in beta version. Reports are disseminated through the usual channels.

- Ph.D. student: perform text mining experiments on Amphibians and Insect data, with special attention to evaluation methods that exceed standard ones (precision, recall, F-score, accuracy, inter-annotator kappa). Investigate mutual dependencies between textual analysis, semantic annotation in STITCH, script analysis in SCRATCH, and image processing in RICH, in provisional integration subproject(s)
- Postdoc: supervision of demonstrator development; assist Ph.D. student in experiments; project supervision
- Scientific. programmer: develop demonstrator; assist in integration; prepare web demos

Year 4

Additional analyses and evaluations are performed across the problem areas studied, focusing on the generalities rather than the peculiarities of each area. The Naturalis software demonstrator is released and demo-ed. The Ph.D. student writes his/her thesis.

- Ph.D. student: continue fundamental experimental work, focusing on evaluation of text mining methods; write thesis
- Postdoc: continue integration & mutual benefits research
- Scientific. programmer: continue demonstrator development; document software; develop web demos; prepare for maintenance

3f) Deliverables

The project aims to deliver the following products of research:

- Software environment demonstrators: First, a demonstrator showcasing the software developed within Naturalis. Second, an extension to this demonstrator illustrating the cooperation with script analysis in the SCRATCH project. Third, provisional demonstrators based on the cooperation with image processing in the RICH project and semantic annotation in the STITCH project.
- Journal articles. Targeted journals: Computing in the Humanities; Natural Language Engineering; Computational Linguistics.
- Conference proceeding papers. Targeted conferences: ACL (Meetings of the Association for Computational Linguistics); COLING (Intl Conference on Computational Linguistics); HLT (Human Language Technology); LREC (Linguistic Resources and Evaluation Conference); ICML / ECML (International / European Conference on Machine Learning), and any appropriate workshop – preferably, satellite events to major conferences or Belgian-Dutch yearly peer-group workshops such as CLIN (Computational Linguistics in the Netherlands), BENELEARN (Belgian-Dutch Machine Learning workshop) and BNAIC (Belgian-Dutch AI Conference).
- Ph.D. thesis.

4) Expected use of instrumentation

The team needs sufficient computing power besides normal desktop computers to operate. One high-end computer (dual-CPU, minimum 2 Gb memory and ample permanent storage capacities in terms of hard disks and DVD RW) will act as computing server. This high-end server will be physically integrated in the Tilburg high-end computing infrastructure (with particular maintenance demands for which the Tilburg group has expertise). Experiments on this machine can be done remotely from the workstations at Naturalis.

5) Literature

5a) References to cited works

Buchholz, S. and Van den Bosch, A. (2000). Integrating seed names and n-grams for a named entity list and classifier. In: *Proceedings of LREC-2000*, Athens, Greece, June 2000, pp. 1215-1221.

Daelemans, W., Buchholz, S., and Veenstra, J. (1999). Memory-based shallow parsing. In: Proceedings of CoNLL-99, Bergen, Norway, June 12, 1999.

Hendrickx, I, and Van den Bosch, A. (2001). Dutch word sense disambiguation: Data and preliminary results. In Proceedings of SENSEVAL-2. Toulouse, France.

Hearst, M. (1999) Untangling text data mining. In Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999

Hendrickx, I., and Van den Bosch, A. (2003). Memory-based one-step named-entity recognition: Effects of seed list features, classifier stacking, and unannotated data. In Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp. 176-179.

Hoste, V., Hendrickx, I., Daelemans, W., and Van den Bosch, A. (2002). Parameter Optimization for Machine-Learning of Word Sense Disambiguation. Natural Language Engineering, Special Issue on Word Sense Disambiguation Systems 8:4, 311-325, 2002.

Jackson, P. and Moulinier, I. (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization. Amsterdam: John Benjamins.

Mitchell, T. (1997). Machine Learning. McGraw-Hil.

Oostdijk, N. & D. Broeder. The Spoken Dutch Corpus and Its Exploitation Environment. In Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03). 14 April, 2003. Budapest, Hungary.

Tjong Kim Sang, E. (2002) Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002, Taipei, Taiwan, 2002, pp. 155-158.

Van den Bosch, A., and Daelemans, W. (1999). Memory-based morphological analysis. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99, University of Maryland, USA, 20-26 July 1999, pp. 285-292.

Van Halteren, H., Zavrel, J., and Daelemans, W. (2001). Improving word class tagging through combination of machine learning systems. Computational Linguistics, 27:2, pp. 199-230.

Zavrel, J. and Daelemans, W. (2003). Feature-Rich Memory-Based Classification for Shallow NLP and Information Extraction. In Jurgen Franke, Gholamreza Nakhaeizadeh and Ingrid Renz (eds.), Text Mining, Theoretical Aspects and Applications, Springer Physica-Verlag, pages 33-54.

5b) Five most important publications of the research team

Daelemans, W., Van den Bosch, A., and Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. Machine Learning, 34, 11-43.

Hoste, V., Hendrickx, I., Daelemans, W., and Van den Bosch, A. (2002). Parameter optimization for machine learning of word sense disambiguation, Natural Language Engineering, 8:4, pp. 311-325.

Van den Bosch, A. (1999). Careful abstraction from instance families in memory-based language learning. Journal of Experimental and Theoretical Artificial Intelligence, 11:3, pp. 339-368.

Van den Bosch, A., and Buchholz, S. (2002). Shallow parsing on the basis of words only: A case study. In Proceedings of the 40th Meeting of the Association for Computational Linguistics, New Brunswick, NJ: ACL, pp. 433-440.

Van Halteren, H., Zavrel, J., Daelemans, W. (2001), Improving accuracy in word class tagging through combination of machine learning systems. Computational Linguistics 27:2, pp. 199-230.

Core Project 640.003.401

1a) Project title
Cultural Heritage Information Personalization

1b) Project acronym
CHIP

1c) Principle investigator
Prof. dr. P.M.E. De Bra (TU Eindhoven)

1d) Main project location
Rijksmuseum

2) Composition of the research team

- 1 Ph.D Student
- 1 Postdoc
- 1 Scientific programmar
- P. Sigmond (Rijksmuseum)
- P. Gorgels (Rijksmuseum)
- P. De Bra (TU/e)
- G.J. Houben (TU/e)
- L. Aroyo (TU/e)
- M. Veenstra (TI)
- R. Brussee (TI)
- M. Alberink (TI)

3) Description of the proposed research

3a) Problem statement and research objectives
The cultural heritage collections and relevant related contextual information are distributed over many different institutes. CATCH aims to make the barriers between these institutes disappear by providing virtual integration of these collections. The CHIP project focuses on the interaction of the users with the combined cultural heritage content, and in particular on personalized presentation and navigation. Cultural heritage information (digitized versions of cultural artifacts as well as descriptive information) is used by a wide variety of user types, ranging from schoolchildren to professional art experts (including museum curators and researchers). There have already been initiatives within the cultural heritage sector to provide information in a variety of formats ranging from websites and leaflets (printed on demand) to audio tours. The CHIP project aims to give attractive presentations of this information in a way that is appropriate for the actual user, and that is presented in a form suiting the user's characteristics, his usage context and his computing and communication equipment. In addition the scope of the CATCH project gives the unique opportunity (and challenge) to make presentations that span several collections and background information from libraries, museums and potentially broadcasters and magazines and newspapers. Through the use of portable devices the personalized information can not only be offered on the Web but also inside museums, using the user's location in addition to all the other information about the user for performing the automatic personalization.

Cultural collections do not consist of a discrete database of art objects. They come with a story that connects different art objects, details of objects, and background information together. Explaining and exposing this story is an important task of the cultural heritage institutions. Moreover finding, cataloguing and explaining such stories is an import part of the work of professionals. Together these stories impose a structure on the collections. In fact they impose many different structures. The CHIP project aims to present art objects by organizing them in groups, and more generally showing the relations between art objects, based on the objects metadata, and indirectly on their links with background information. Hypermedia formats will be used to allow multi-branched story lines that can connect art objects in “precooked” or automatically generated ways. Adapting the way objects are organized is also an import aspect of personalization as the usefulness of structure largely depends on the knowledge of the user and the medium that is used. A challenge will be to find ways to achieve this goal while leaving responsibility for the collection to different institutions including some form of responsibility for their presentations.

Closely related to imposing structure is the problem of navigating through the heterogeneous collections of the institutions. CHIP does not want to restrict the user to follow a predefined path, and so a navigation structure is needed to prevent the user from getting lost and to provide information smell. A combination of searching and browsing is envisioned. Especially visually oriented browsing will be considered as an integral part of the presentation.

Personalization is done in two stages: first through the definition of stereotype user groups, and later through adaptation to individual user characteristics. A first approach to gathering information about the user is by explicitly asking the user about his/her intentions and preferences through a (short) questionnaire. This allows for the creation of adaptable sites. Unfortunately this has the drawback of forcing the user to perform some actions before gaining access to the cultural heritage information. By presenting a few links like “information for...” the fact that stereotype user modeling is being performed can be partially disguised. A second approach is to deduce user information from observing the user’s browsing and searching behavior. Continuous observation results in adaptive sites, which have the advantage that the site’s behavior adapts itself to changes in the user’s behavior. Adaptivity is more difficult to achieve and has the drawback that when a new user first uses an adaptive site, the site has no information to start with and will have to start with a lowest common denominator presentation. Thus a combination of adaptable and adaptive features appears to be in order: asking the user one or more simple “how can I help you” type of questions or presenting “information for...” links to identify a stereotype such as child, tourist, researcher, and then adapting as the user moves along.

Personalization requires user profiles to be constructed. Cultural heritage institutions are interested in being able to reuse these profiles to better understand their visitors, both on the web and when physically visiting the museum or institute. Hence, managing these profiles in a way that allows the extraction of such information is an important side issue. Of course, in order to do meaningful adaptation user information needs to be gathered from the browsing and searching in the combined information from the virtually integrated cultural heritage sites. There are essentially two ways to realize the required user profiling: by redirecting the access to the information through a “proxy-like” gateway, or by coupling the different servers to a joint user modeling service. We actually propose a combined approach, with an application server taking care of the more global user goals and the development of a global presentation strategy, an adaptation server taking care of the more localized adaptation and a user model server taking care of storing and interpreting user-related information. This architecture is currently being investigated in the Token2000 project CHIME. The CHIP project will perform validation of that model by trying to use it to create the architecture to provide personalized access to the combined cultural heritage information.

3b) Research questions and approach:

The CHIP project is concerned with presentation, navigation and personalization. Although these aspects are of course related we try to separate them here:

- **Presentation:** this deals mainly with the problem of presenting cultural information in all its diversity; in particular we concentrate on the following issues:
    - A side-effect of the virtual integration of different databases is that different information is available for different objects (because they come from differently structured databases). Experience from the Topia project of TI, CWI, IBM and TU/e shows that such structural differences are easier to handle when the used information model is closely related to the semantics, rather than using the data model of the underlying database. Work on semantic integration will provide such a layer. Existing approaches for presenting structured information (e.g. the Hera project at the TU/e)

    work well for uniform structures, but need to be extended to cover the generation of presentations of information objects of differing structure. The Hera and Topia approach will be the starting point for researching this problem [Barna04, Houben04].
  - A second aspect is that although CATCH aims at virtual integration of collections each cultural heritage source may wish to be "identified" in the presentation. This is actually less of a technical challenge than of a political and organizational one. A negotiation will be needed between the different cultural institutes to find a compromise between the uniformity of the presentation of cultural information and the desire to show the particularities of the specific collection and a reference to the source or owner of the information.
  - CATCH will provide the integration of information about cultural objects from different sources. However, this does not automatically result in a sensible "story" [Rutledge03]. The problem of combining different information fragments into a sequence (or other structure) is known as narrative smoothing. In the CHIP project we will investigate narrative smoothing, but concentrate on what information must be provided by the different information sources in order to ensure that a sequence of information fragments is sensible. (A simple example of this is the provision of dates in order to create a chronological sequence.) Narrative smoothing at the sentence level is subject for further research, beyond the scope of CHIP. (It also requires expertise of computer linguists, not involved in CHIP.) In CHIP the integration approach will be based on the existing collaboration between the VU, CWI and TU/e [Stuckenschmidt04, Rutledge04].
- **Navigation:** Traditionally navigation in hypermedia has been considered to be just browsing (or following links). However, in the wealth of cultural information it is impossible to find and study all the information a user needs by just browsing. Therefore a combination of browsing and searching needs to be developed that enables users to quickly get to the desired information. A first study of browsing+searching has been performed by Aroyo [Aroyo01, Aroyo02].
  - We will consider a two (or more) level hypermedia architecture (like originally developed by Bruza [Bruza90]) to move from content level to concept level, search or browse at the concept level and then "beam down" to the content level again. The navigation through the found information can be generated from the underlying structure. Methods derived from the ideas of RMM [Isakowitz95] translate database relationships into navigational relationships.
  - Searching and browsing need to be seamlessly integrated. Based on the possible number of choices a presentation generator may decide to show links (when there are few choices) or a search form (hiding the fact that there are only a fixed but large number of choices). The choice between searching and browsing is also subject to personalization. Experts may prefer search interfaces that let them specify exactly what they are looking for whereas casual visitors (or school children) want closed sets of choices, presented through a browsing interface. We will create a number of navigation and search interfaces to perform usability testing on in order to determine the most suitable way to reach desired information for different user categories.
- **Personalization:** The biggest challenge for on-line presentation of cultural heritage information is the adaptation to the widely varying audience. We divide the issues of personalization into adaptable and adaptive aspects, and also study the issue of user modeling:
  - Personalization always depends on user characteristics. These are typically stored in a user model that can be initialized based on a simple questionnaire (or a set of links) to identify a stereotype, and then updated by tracking the user's browsing and searching behaviour. The biggest user modeling issue for CHIP is that there are multiple stakeholders. Different sources of information may wish to gather information about the users, whereas a unified portal (envisioned by CHIP) prefers to do all the user modeling by itself. The challenge is to provide user-related information to all the partners without redundant storage (and without violating privacy laws, but this is a side issue left for future research projects). In order to tackle the technical challenge we will develop a modular architecture of communicating user model servers. (We are currently investigating the user model server architecture and UserML language developed at the University of Saarbrücken [Heckmann03]. The main technical problem is to ensure good performance, whereas the scientific

challenge is to ensure that a proper translation is made (if needed) between the terminology or concepts used in the individual user models.

- The challenge regarding making a combined presentation adaptable is that the different sources may offer information for a different partitioning of the user population. This means that they may consider different stereotypes, that must be mostly mapped to each other. Another problem is that a stereotype must be identified quickly as users do not wish to spend a lot of time answering questions before getting to the actual information a site has to offer. In order to find out the difficulty of the problem of determining the stereotypes we will investigate the target user groups for different information sources that are combined. In CATCH as a whole this problem will be more difficult than in the environment studied in CHIP which does not cover all of the Dutch cultural heritage. Stereotypes also relate to different aspects of users, so a combination of smaller aspects may be needed. There are issues related to the user's knowledge and attitude, for instance distinguishing school children from adults and from art experts. But there are also issues related to the device and network that is used, the amount of time available or specific disabilities. The adaptation to a stereotype may influence both the content of the presentation and the navigation. A generated audio tour will be more sequential than a website for instance, so that influences navigation, whereas the difference between a school child and an art expert will be first and foremost related to the content that is presented.
- There are endless ways to navigate through a richly structured site. What makes the experience truly personalized is the adaptive behavior of a site. The content can refer to information the user saw earlier, and the suggested links may depend on the perceived interest of the user. Contrary to typical adaptive hypermedia systems (as described in [Brusilovsky96, Brusilovsky01], the adaptation cannot be completely foreseen by an author, as the information is gathered from databases that are not under the control of a single author. Therefore the adaptation needs to be defined at a higher, conceptual or schema level. Information retrieval techniques for semantic information will be investigated for this part. Related information can be recognized because it will be associated with the same concepts and this is stored in metadata. Typical adaptive hypermedia techniques like the use of prerequisites to perform link adaptation or sorting need to be based on prerequisites that can be automatically deduced from metadata. For instance, when the user wishes to view (information about) objects in chronological order, the dates can be used to generate prerequisite relationships that will result in a chronological presentation. But similarly presentations can be ordered based on sequences of art movements, and the order of presentation of media types (like images before text) can be generated as well. We will investigate how the Hera approach can be used to generate adaptive structures that feed into adaptive systems like AHA! [DeBra03] for delivery of the content to the end-users. This will require extensions to AHA! to enable it to accept the input of new structures at runtime. Also, AHA! will be modularized to make it work with external user model services. The creation of an adaptive delivery platform for CHIP will require substantial development effort, which is why the project proposal includes the provision for a full-time software developer in addition to the research staff (phd student and post-doc). Also, the research and development team must be positioned inside the Rijksmuseum to ensure a close match between what the team develops and the needs of the cultural heritage world and its information technology infrastructure.

3c) Scientific relevance

The CHIP project builds on the existing work in adaptive hypermedia (exemplified by the AHA! system [DeBra03] and the AHAM reference model [DeBra99]) and on dynamic Web-based information systems (exemplified by the HERA research [Barna04, Houben04]). The scientific challenge for CHIP is to combine both types of approaches, and define and study a generic architecture for adaptive web-based information systems (just like AHAM defined a generic architecture for authored adaptive hypermedia applications). Adaptation is typically based on manually authored *concept relationships*. This only works well for reasonably small set of information items (pages). In order to provide adaptation (or automated personalization) in huge information spaces these concept relationships must be deduced automatically from metadata associated with the information items. Also, the adaptation must be based not only on *subject* information but also on *purpose* information in order to perform adaptation not only to individual browsing and searching patterns but also to different target user groups and the tasks they perform. This builds upon research in the CHIME (Token2000) project.

3d) Related Work

In the Topia project, research has been done on the navigation through visual cultural heritage information using presentations dynamically structured based on existing and external meta data and preferences of the user. This collaboration between TI, CWI, IBM and TU/e can be seen as a preliminary step towards starting the CHIP research. Also, the CHIME (Token2000) project in which the VU, CWI and TU/e collaborate aims at an initial architecture to allow task-oriented access to cultural heritage information. The NWO DYNAMO project in which CWI and TU/e collaborated and a related part of the RTIPA ITEA project have resulted in an approach to adaptation to device characteristics. Combining browsing and searching, and presenting information as well as visualizing the corresponding conceptual information is being investigated in the SWALE project (NWO sponsored collaboration with the University of Leeds). The CHIP project is a real challenge to bring isolated approaches to different aspects of the creation of adaptive applications together and produce a distributed adaptive web-based architecture for the cultural heritage field that actually works.

3e) Work programme

The CHIP research team will be stationed mainly at the Rijksmuseum. This museum has a huge heterogeneous database with information about its art collection. CHIP can use this as an abstraction from the virtually integrated cultural heritage collection that CATCH aims at as a whole. A close collaboration with the IT specialists of the Rijksmuseum will ensure that the generic architecture to be developed in CHIP will fit the actual information infrastructure of the Rijksmuseum. The (preliminary) planning of the CHIP project is as follows:

Year 1

The team investigates the literature on adaptive hypermedia and web-based information systems and studies the existing software prototypes that solve parts of the problem (including results from the AHA!, HERA, CHIME, AIMS and SWALE projects). Based on these investigations a first architecture will be designed and implemented, by reusing components and developing communication between the components. The team will also collaborate with researchers of AHA!, HERA, CHIME and other projects in determining ways to make the components more generic and reusable. The division of tasks is roughly:

- Ph.D. student: studying the literature and designing the architecture.
- Postdoc: studying the available information (database) and how it can be translated or linked to existing adaptive or web-based information system components; validation of the match between the architecture being designed by the Ph.D. student and the information infrastructure of the Rijksmuseum.
- Programmer: creation of the interfaces between the components that are borrowed from other projects, and implementation of the designed new architecture by reusing these components.

Year 2

In the second phase of the project the focus will be on the definition and implementation of the desired personalization for cultural heritage information and the different intended audiences or users of that information.

- The post-doc will investigate the available metadata and the used ontologies from which concept relationships can be generated automatically; (s)he will also investigate the types of users of the Rijksmuseum database, and the adaptation that is required for these users and their tasks.
- The Ph.D. student will first investigate the automatic classification of users into user groups and the identification of the users' tasks, in order to determine the basic parameters that guide the initial adaptation; (s)he will then investigate the adaptation of information to these users, and in particular the adaptation in a combined browsing and searching interface, where visualization of concept structures and of pointers (links) to information items, and the items themselves, need to be seamlessly combined.
- The programmer will design and implement the software for the user group detection and

will create the browsing and searching environment, based primarily on the designs that come from the AIMS and SWALE projects.

Year 3

The CHIP architecture, using the Rijksmuseum information, will be tried with users in order to evaluate the design and implementation and to deduce necessary improvements. The post-doc will coordinate the evaluation and the analysis of the resulting information. The Ph.D. student will concentrate on the user-interface design aspects and develop interfaces for use on the Internet (using normal computers with Web browsers) and for use with PDAs inside the museum (using location-based information to augment what the user sees in the museum with information about the artifacts and information about related objects located elsewhere in the museum or in other institutes. The programmer will develop the software for the evaluation and perform the incremental updates to the software and the interfaces.

Year 4

Most of the final phase of the CHIP project will concentrate on the dissemination of the results, first and foremost to other cultural heritage institutes in the Netherlands, but also to the international scientific community. The Ph.D. student will write the final dissertation, and the programmer will finalize the software and documentation and design and implement gateways with databases from other institutes.

3f) Deliverables

The project aims to deliver the following products of research:

- A generic architecture for the adaptive delivery of information from a huge collection of richly (semantically) annotated information items. The architecture will be largely independent of the application in the cultural heritage domain.
- A software platform for automatic personalization of the interaction with cultural heritage information. The interaction is a combination of browsing and searching. The adaptation is performed using initial stereotypes (user groups) and subsequent tracking of individual user goals, tasks and actions.
- An adaptive information system serving Rijksmuseum information. The system will also feature gateways to external databases serving information from other institutes.
- Conference and journal articles, mostly in the area of (adaptive) hypermedia, advanced visual interfaces, user modeling and web engineering.
- A Ph.D. thesis.

4) Expected use of instrumentation

The team needs one medium-size server (in addition to normal workstations) to store the Rijksmuseum database and generate presentations on-the-fly. (dual cpu, minimum of 2GB main memory and disk raid array of 1TB or more) Because the project starts with the reuse of components the initial implementation will not be optimal and therefore need more processing power than the final software that will be produced in year 4 of the project. The final software will be used with larger number of users, which will still make the use of powerful hardware necessary. It is expected that the real deployment of the CHIP architecture will require server farms, but these are not necessary during the research phase.

5) Literature

5a) References to cited works

[Aroyo01] Aroyo, L., Task-oriented Approach to Information Handling Support within Web-based Education, Ph.D. thesis, University of Twente, 2001.

[Aroyo02] Aroyo, L, Dicheva, D., AIMS: Learning and Teaching Support for WWW-based Education, International Journal for Continuing Engineering Education and Life-long Learning (IJCEELL), 11, No. 1/2, pp. 152-164, 2004.

[Barna04] Barna, P., Houben, G.J., Frasincar, F., Specification of Adaptive Behavior Using a General-Purpose Design Methodology for Dynamic Web Applications, AH2004, Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Eindhoven, the Netherlands, Springer LNCS 3137, pp. 283-287, 2004.

[Bruilovsky96] Brusilovsky, P., Methods and techniques of adaptive hypermedia, UMUAI, Journal on User Modeling and User-Adapted Interaction, 6(2-3), Kluwer, pp. 87-129, 1996.

[Brusilovsky01] Brusilovsky, P., Adaptive hypermedia, UMUAI, Journal on User Modeling and User-Adapted Interaction, 11(1-2), Kluwer, pp. 87-110, 2001.

[Bruza90] Bruza, P., Van der Weide, Th., Two Level Hypermedia – An Improved Architecture for Hypertext. DEXA90, Database and Expert System Applications Conference, Vienna, Austria, pp. 76-83, 1990.

[DeBra03] De Bra, P., Aerts, A., Berden, B., De Lange, B., Rousseau, B., Santic, T., Smits, D., Stash, N., AHA! The Adaptive Hypermedia Architecture, HT'04, Fourteenth ACM Conference on Hypertext and Hypermedia, Nottingham, UK, pp. 81-84, 2003.

[DeBra99] De Bra, P., Houben, G.J., Wu, H., AHAM: A Dexter-based Reference Model for Adaptive Hypermedia, HT'99, Ninth ACM Conference on Hypertext and Hypermedia, Darmstadt, Germany, pp. 147-156, 1999.

[Heckmann03] Heckmann, D., Krueger, A., A User Modeling Markup Language (UserML) for Ubiquitous Computing, UM2003, Ninth International Conference on User Modeling, Johnstown, USA, pp. 393-397, 2003.

[Houben04] Houben, G.J., Frasincar, F., Barna, P., Vdovjak, R., Modeling User Input and Hypermedia Dynamics in Hera, ICWE2004, International Conference on Web Engineering, Munchen, Germany, Springer LNCS3140, pp. 60-73, 2004.

[Isakowitz95] Isakowitz, T., Stohr, E., Balasubramanian, P., RMM: a methodology for structured hypermedia design, Communications of the ACM, 38(8), pp. 34-44, 1995.

[Stuckenschmidt04] Stuckenschmidt, H., Vdovjak, R., Houben, G.J., Broekstra, J., Index Structures and Algorithms for Querying Distributed RDF Repositories, WWW2004, The Thirteenth International World Wide Web Conference, New York, USA, pp. 631-639, 2004.

[Rutledge03] Rutledge, L., Alberink, M., Brussee, R., Pokraev, S., Van Dieten, W., Veenstra, M., Finding the Story – Broader Applicability of Semantics and Discourse for Hypermedia Generation, HT'03, Fourteenth ACM Conference on Hypertext and Hypermedia, Nottingham, UK, pp. 67-76, 2003.

[Rutledge04] Rutledge, L., Houben, G.J., Frasincar, F., Combining Generality and Specificity in Generating Hypermedia Interfaces for Semantically Annotated Repositories, Interaction Design and the Semantic Web, Workshop at WWW2004, The Thirteenth International World Wide Web Conference, New York, USA, 2004.

5b) Five most important publications of the research team

Aroyo, L., Dicheva D., AIMS: Learning and Teaching Support for WWW-based Education, International Journal for Continuing Engineering Education and Life-long Learning (IJCEELL), Vol. 11, No. 1/2, pp. 152-164, 2001.

Aroyo, L., De Bra, P., Houben, G.J., Vdovjak, R., Embedding Information Retrieval in Adaptive Hypermedia: IR meets AHA!, The New Review of Hypermedia and Multimedia, Vol. 10, (page numbers pending) 2004. Taylor and Francis Group Publishers.

De Bra, P., Brusilovsky, P., Houben, G.J., Adaptive Hypermedia, From Systems to Framework, ACM Computing Surveys, Symposium Edition, 1999. Publisher's reference: http://www.acm.org/pubs/articles/journals/surveys/1999-31-4es/a12-de_bra/a12-de_bra.pdf.

Rutledge, L., Alberink, M., Brussee, R., Pokraev, S., Van Dieten, W., Veenstra, M., Finding the Story – Broader Applicability of Semantics and Discourse for Hypermedia Generation, HT'03, Fourteenth ACM Conference on Hypertext and Hypermedia, Nottingham, UK, pp. 67-76, 2003.

R. Vdovjak, F. Frasincar, G.J. Houben, P. Barna, Engineering Semantic Web Information Systems in Hera, in: Journal of Web Engineering, Vol. 2, No. 1/2, p. 3-26, 2003, Rinton Press.

APPENDIX II: Involvement of consortium members in related international research projects

Prof. dr. Jaap van den Herik, Universiteit Maastricht/Universiteit Leiden

- International Network for the Conservation of Contemporary Art: Science and Technology (FP6: Integrated Network Proposal)
- Kdnet: European Knowledge Discovery Network of Excellence (FP5-IST project)
- CEPIS: a European non-profit organisation seeking to improve and promote high standards among informatics professionals in recognition of the impact that informatics has on employment, business and society
- EUROPAC: EU-compliance Regulatory Ontologies Platform for Assurance and Certification
- EMMI - Euregional Multi Media Information exchange (PF4-Telematics project)

Drs. Paul Doorenbosch, Koninklijke Bibliotheek

- Het Geheugen van Nederland: national project involving more than 35 Dutch cultural heritage institutes with international extensions, a.o. with the Library of Congress, The New York Public Library and the British Library

Prof. dr. Frank van Harmelen, Vrije Universiteit

- On-To-Knowledge, early project on Semantic Web technology (FP5-IST project)
- IBROW, early project on Semantic Web services (FP5-IST/FET project)
- Wonderweb, Project on Semantic Web technological infrastructure (FP5-IST/FET project)
- SWAP, Project on exploiting peer-to-peer technology for Semantic Web (FP5-IST/FET project)
- OntoWeb, Network combining all ontology-related research in Europe (FP5-IST project)

Prof. dr. Paul De Bra, Technische Universiteit Eindhoven

- ADAPT Adaptivity and adaptability in ODL based on ICT Minerva project with the University of Twente, University of Southampton, University of Nottingham, Politecnico di Milano and IRST, Trento.
- AHA!, Adaptive Hypermedia for All, project funded by the NLnet Foundation, with collaboration of the University of Pittsburgh and the University of Southampton.
- PROLEARN, Network of Excellence in Professional Learning (FP6-IST Network of Excellence. TU/e is one of 19 core partners).

Prof. dr. L. Schomaker, Rijksuniversiteit Groningen

- WANDA, Forensic writer indentification (German-Dutch project)
- MIAMI, Multimodal Interaction in Advanced Multimedia Interfaces (ESPRIT-BRA project)
- International UNIPEN Foundation, a project for the collection of hand-written databases and the benchmarking of handwriting-recognition systems

Dr. Antal van den Bosch, Universiteit van Tilburg

- MUSA Multilingual subtitling of Multimedia Content (PF5-IST project)

Prof. dr. ir. Chris Vissers, Telematica Instituut

- ENRICH: ENhanced authoRIng and design of advanCed multimedia content tHrough introduction of the pantheon methodology and the integrated asset factory (FP5-IST PPP)
- COCONET: COntext-Aware COllaborative Environments for Next Generation Business NETworks
- MINDS, Multimodal interaction for natural dialog systems (FP6 project proposal)
- CREATE, Knowledge Supported Service-Based Cooperation Workspaces Enabling People in Networked Organisations to Create Business Opportunities and Share Resources (FP6 Integrated Project proposal)

Dr. P. Wittenburg, Max-Planck-Institut für Psycholinguistik, Nijmegen

- ECHO: European Cultural Heritage Online (FP5 IST project)
- INTERA: Integrated European Language Resources Area (EU E-Content project)
- CHASE: Web for Culture, History and Science for Europe (FP6 Integrated Project proposal)
- Member ISO/TC37/SC4: Language Resources Management