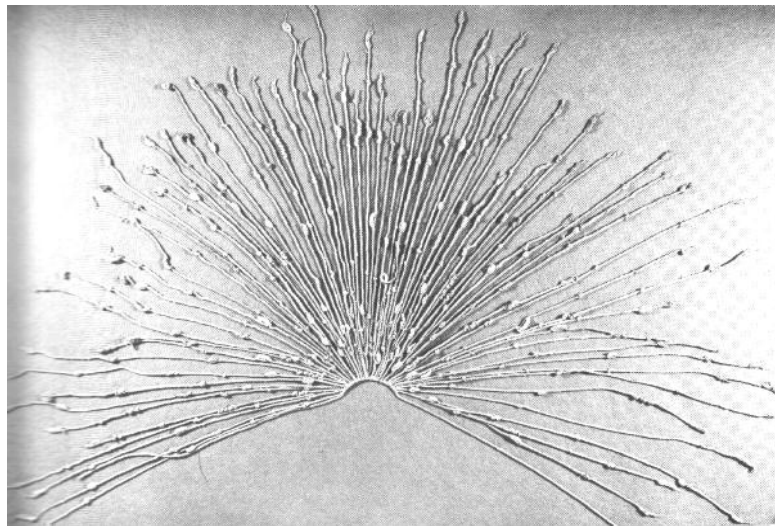


E-based Humanities and E-humanities on a SURF platform

Joost Kircz



KRA *publishing research*

Cover picture: A quipu from Peru (source: Albertine Gaur. A history of writing. The British Library, 1992)

A quipu is an assemblage of coloured knotted cotton cords. The colours of the cords, the way the cords are connected together, the relative placement of the cords, the spaces between the cords, the types of knots on the individual cords, and the relative placement of the knots are all part of the logical-numerical recording.

Quipus are an elementary example of a digital information system.

Most of the quipus are destroyed by the Spanish colonialists, providing an excellent warning that the preservation of knowledge in coded information is not only a matter of erosion and physical destruction of the carrier material, but also highly dependent on the cultural capacity to describe coding schemes in a univocal way that permit transformation into new media.



The "Indian Chronicler" Felipe Guaman Poma de Ayala, writing toward 1613, drew an encounter at a "Collca" or "Warehouse of the Inca": Tupac Inca Yupanqui (left) interviews his accountant or warehousekeeper (right). The warehousekeeper is extending a cord record or quipu, which contains records of goods in the storage chambers. Source: (Geometry step by step from the land of the incas. http://agutie.homestead.com/files/Quipu_B.htm)

E-based Humanities and E-humanities on a SURF platform

A report commissioned by SURF-DARE

Author: Dr Joost Kircz

Kircz Research Amsterdam

www.kra.nl

1 June 2004

E-based Humanities and E-humanities on a SURF platform

Contents:

- 1- Introduction
- 2- This report
- 3- What is E-science
 - 3.1- Nomen est Omen
 - 3.2- E-based versus E
- 4- Humanities versus natural science
 - 4.1- Intrinsic or extrinsic cultural differences?
 - 4.2- Quantitative studies as a rule?
- 5- Five levels of processing
 - 5.1- Defining the information object
 - 5.1.1- Naming entities
 - 5.1.2- Descriptive languages
 - 5.1.3- Indices and ontologies
 - 5.2-. Writing, rewriting and presenting
 - 5.2.1- Rewriting of existing and new textual material
 - 5.2.2- Image digitisation
 - 5.2.3- 3D visualisation
 - 5.2.4- Sound and film
 - 5.2.5- Thinking
 - 5.3- Storing information
 - 5.3.1- The digital archive
 - 5.3.2- Dedicated storage environments
 - 5.3.3- Chaining
 - 5.4- Analysis of digital corpora
 - 5.4.1- Two poles
 - 5.4.2- Multimedia tools
 - 5.5- The reading and using
- 6- Conclusions

Acknowledgement

Glossary

References

- (R1). Scientific literature
- {R2}. Report literature
- [r3]. Websites

List of interlocutors

Note: The references are split in three categories.

(x) references to scholarly publications

{y} references to report literature

[z] references to websites

1- Introduction

As of 2003, SURF[0] enables three platforms: ICT and Research, Education, and Organisation. Within these programmes, SURF has funds available to promote ICT innovations. Innovation is not an easy notion to explain. Too often we encounter new wine in old bottles and changes in vocabulary frequently cover the continuation of existing programmes in a new framework.

This explorative study was commissioned to review the underlying issues and to suggest new directions for research that will exploit ICT in the Humanities to our best advantage. Thus, the purpose of this study is to provide an insight in the complicated relationship between so-called E-techniques and the humanities or better E-humanities. Based on this general insight, suggestions for a SURF policy in the E-humanities are offered.

2- This Report

The main goal of this study is to reach an understanding of to what extent SURF can enhance, develop and change the way and rate at which humanities research develops in an electronic environment. An important question is also to what extent we can identify common problems, challenges and outlooks?

As SURF is a Stimulating, Initiating and Facilitating organisation, the emphasis of this indicative study is to clarify the particularities of ICT in the humanities vis-a-vis the natural sciences in order to suggest directions that will enable SURF to fulfill its mission in the broadest possible way.

Given the short time frame of two months allocated for this study, the emphasis is on those fields in the humanities where ICT applications are not yet as common as in the natural sciences. For a general Technology Assessment study of ICT in science, I refer to the excellent book of Michael Nentwich: *Cyberscience. Research in the age of the Internet.* (Nentwich03). In general, one can say that disciplines where mathematical modelling, statistical analysis and algorithmic thinking are common, such as parts of economics, psychology, political science and sociology, the lead of the natural sciences is followed. In less quantitative (sub) disciplines we see serious thresholds in applying “electronic” methods.

The present report is an inductive study, based on literature study and a series of discussions with key scientists and programme directors of relevant institutions.

It was a general concern, in the literature and with the interviewees, that the ICT aficionados deliver solutions whilst the underlying scientific quests are not always explicated for benefit of the humanity scholars themselves. The desire to elucidate the complex situation, in particular for qualitative studies, was a common feature in all conversations. Hence, the main thrust of the report is to clarify the underlying problems.

Without any claim for completeness, the cross section of disciplines in the humanities that were touched upon clearly shows that, on the abstract level of ICT architecture, many common issues can be identified, or to use a phrase of McCarty, “*Methodological Commons*” (McCarty03) can be identified.

The report starts with a discussion on the definition of E-science followed by a comparison between the natural sciences and the humanities. Then, the report is divided in sections dealing with issues instead of sections that analyse different disciplines as such. In particular the distinction between general applicable qualitative methods versus quantitative and not yet fully quantitative approaches is a thorough going thread.

In the last part of the report, I wrap up the lessons learned, and give suggestions for a SURF policy in the humanities.

In order not to repeat discussions, as far as possible previous reports and studies are used. The list of references is split into one part dealing with scholarly publications, one part with report literature and a third one with relevant websites. The appendix also contains an acknowledgment, a glossary, and a list of all interlocutors.

3- What is E-science?

Every new stage in technology creates its own special jargon and hence every new phase in research tends to incorporate a new vocabulary. For that reason, it is important to clearly define notions and processes, as too easy, highly ideological and fashionable metaphors hide the underlying quest. This is all the more important as certain very useful but typical technical concepts have strong rhetorical power and tend to obscure the discussion. The reader will probably recognise such catch-phrases as: search-engine, electronic-highway, intelligent-systems, knowledge-mapping, grids, virtual environments, data farms, etc.

However, the *real* issue is, whether and how such technological developments can and will change the way research is done, and, vice versa, how scientific quests induce new technological developments. This issue lies at the heart of what E-science is, or better, ought to be about.

This dialectic cannot be cast in terms of “question and demand”, “push and pull” or “market forces” as such rhetoric tries to contain the real problems into a known trivial linear grammar.

3.1- Nomen est omen

In the same way as in a previous phase, the term Multi-Media became an epithet for everything that was hot; now the term E-science is used that way.

Nentwich (Nentwich03, p21) discusses the use of prefixes like e-, tele- and cyber- . In fact the e- prefix is normally used for everything that uses electronic means, the opposite to older forms of empowerment, and tele- just means everything farther than arm’s length. The term Cyberspace is clearly coined as “*the virtual space created by electronic networks*”. Hence Nentwich opts for the term Cyberscience as the science and research falling under cyber conditions, which -in essence- means science in an environment that is intrinsically geographically nonlocal and where multiple use and re-use of knowledge objects are the normal situation. In other words, a distributed environment where digital corpora are used simultaneously by different users.

A lot of ICT in the humanities, as we will see below, fits this description. However, not all ICT usage is deemed to be nonlocal. For that reason, in this report I prefer not to use the term Cyber-Humanities.

The latest addition in the list of prefixes is the term Grid. For a clear overview of this notion of a distributed computer infrastructure that allows coordinated use, I refer to Nikolai Petkov's review "Grid computing en e-science" in "De Vruchten Plukken" {Plugge03, pp77-101}. Grid computing is defined as the realm of many Teraflops calculations and Tera-to Petabytes storage and bandwidth. It goes without saying that the need for this type of resources comes from fields where enormously different qualities of data are generated and / or need to be analysed quantitatively (Hey03). We have to be careful not to equate cyberscience or E-science with Grid computing. It is also important to warn the reader that nowadays the term Grid is also sloppily used for all kinds of interconnected collaborations.

To illustrate the rather limited view of marketing the Grid, we quote from *The National e-Science Centre Report 2001-2002* of the UK national programme {NeS02, pg 5}. "*E-science will provide technology which can transform research in any discipline. By combining the expertise of the world's leading experientialists, theoreticians and computer scientists, we will develop distributed computing systems that are capable of storing and analysing the ever expanding volumes of data produces by today's scientific researchers. Extensive global collaborations will become a possibility, advancing the capabilities of computing, communication and visualisation*".

We see that here, E-science is restricted to a technological aid for distributed storage, retrieval and communications.

The *UK e-Science Core Programme Annual Report* {Uke02} aims are described as: "*to research, develop and implement key features of a communication, computational and data (Grid) infrastructure, which will support scientists engaged on UK "grand challenge" science projects*".

So, what is left over for those disciplines that are still struggling with the quest to what extent can electronic means assist in qualitative research and to what extent can quantitative techniques be developed and used in their domain? The methodological challenge is often overlooked.

3.2- E-based versus E

In order to delineate the uneven phases in the development, in this report we make a clear distinction between two essentially different phases of "E-science". On the lower level, we have the use of E-services as enhancements of current practice. Here we deal with the creation of electronic corpora and the development of qualitative methods applied to electronically available data sets. On this level, I also count the new established communicative functions such as e-mail, file transfer, web-interfaces, regular internet usages, and digital libraries. In other words: ICT is used as a supporting tool enhancing established methodologies and techniques. In the following I call this E-based humanities.

On the higher level, we deal with those applications and research quests that only become addressed in a fully digital environment. In the following, this will be named E-Humanities.

If people are confronted with clearly structured electronic files and at the same time have proper authoring tools, the acceptance and understanding of electronically-based science will become obvious. This is simply a chicken and egg problem. In evolutionary biology the egg was first, as the genetic changes that let to the evolution of a chicken took place in the sex

cells and came together in the egg on fertilisation, which meant that the first chicken must have hatched from an egg laid by a bird that was not a chicken.

In the same way, in science, it are the mutated knowledge representations, expressed in a digital form, that merge in the digital repository, therewith enabling the hatching out of novel forms of science. In the present intermediate stage of digitisation of existing information and knowledge, the emphasis is still on the egg (container) side. The ingredients are now prepared and mixed. The objective of the SURF programme has to be that this mixture enables the next phase in scholarly research.

As will become clear in the following, E-science in, e.g., computer science, where totally novel concepts are worked out, can provide the E-based preconditions for E- Humanities.

4- Humanities versus Natural Science

4.1- Intrinsic or extrinsic Cultural differences?

This report is not the place to expand in length on the differences of “*The two Cultures*” and the question to what extent we need to enhance the scientific literacy of non- natural scientists or vice versa (Snow59). However, the different approaches, which are deeply rooted in the traditions of the various disciplines, form an important ingredient for a better understanding of the possibilities of an E-humanities programme. In almost all discussions, the distinction in approach was mentioned. A fair share of the review *Past, Present and Future of Historical Information Science* of Boonstra, Breure and Doorn (Boonstra04) also deals with this issue. Humanities’ scholars are not used to thinking in terms of creating their own technological environment. They take the given environment for granted and hope that it will be useful for them. A bit exaggeratedly, one can say that where the natural scientists can be typified as those people who like to look under the hood of the car and start to take equipment apart before reading the manual, it is the humanities scholar who wants to drive the car without even the need for reading the manual. In the past decade, a series of most interesting projects in so-called *alpha-informatics* has been carried out, without much spin-off. A pessimistic interviewee even went so far as to blame the low success rate of previous projects completely to the cultural differences, where humanists are rated by their peers based on the books they publish instead of on the (inter-) operability (or re-usability) of their production. A (art)book full of images is still preferred above a well-indexed CD-rom or on-line product that enables dynamic comparison between images and manipulation of the metadata in order to analyse the collection or corpus. Another scholar blamed the enormous success of office automation, which its intrinsic simplistic logic, as one of the causes of the present difficulties. A one-size-fits-all dream that unfortunately has no relationship with reality. Scientific calculations are then reduced to office applications.

Thus, we are facing the serious question, why so many examples and exemplars that emerged from application projects of ICT tools in the humanities created neither great enthusiasm nor following. A lot of pilot projects are good show-cases of what could be done but remained “products without a buyer”. An example is the series of CD-roms on Paintings with Iconclass [1] indexing (e.g., Courtauld98).

On the one hand it is of course, as mentioned above, the result of overselling the merits of those methods that are based on first order logic and therefore ideal for an office automation

approach. The whole ICT and Internet hype is geared to the fallacy that “*ICT is doing it for you*” which only enhances the already existing feeling that there must be some male person out there that can fix the problem or at least explain why I’m suddenly forced to use MS-Outlook instead of Eudora as mail program (a concrete problem mentioned). The issue here is, that it is not clear when for standardisation of administrative purposes a certain selection is made and when in a research environment dedicated tools are needed. For instance, in fields that use a lot of maths, LaTeX as a text processor is still the preferred choice, despite the fact that it is difficult to translate in XML.

On the other hand, it goes without saying that at the first educational level, in the whole field of humanities, almost nothing is being taught on methods and techniques that force the young researcher to properly define the object of study in analytical terms. Students get lessons in computer usage, but these are more training courses on ready-made applications instead of courses to understand how problems have to be defined and how features and relations between features can be named and manipulated. This tendency is already emphasised in earlier studies {Voorbij97}. In short, for many linguists and historians, ICT is reduced to formal data manipulation, storage and communications and is not seen as an area where intrinsic research quests can be explicated and dealt with: only a small number of very active trailblazers prove that very deep and exiting research is possible. The importance of such trailblazers for success is also a conclusion of Nentwich (Nentwich03, pp175-181).

A third important intrinsic difference is that in the natural sciences, controlled experiments are based on a theoretical model and given the state of the art of technology. The experimental design of the Large Hadron Collider (LHC) at CERN, or a satellite for astrophysical measurements, forces the soft- and hardware developers into the realm of terabyte or even petabyte manipulation and Gridcomputing. The same is true for environmental sciences and the many “omics” (e.g., Genomics, Proteomics, etc.) in the biomedical world (Hey03).

In contrast, in the humanities, it is the already existing sources that dictate the reach of research. These sources are a given and not a result of modelled experimentation. Having said that, immediately we have to stress that this is only partly, historical, truth. The real, physical, objects of investigation in the natural sciences as well as in the humanities are what they are, a sugar crystal, an incunabulum or pangs of love. The (scientific) representation model tries to come as close as possible to material reality. However, in the natural sciences, the building of representations of the objects of research, such as the collection of certain spectrographic data, is much further advanced. In the natural sciences, we *create* a research corpus in the measurement. Presently, in the humanities, we are only starting to *recreate* the research corpus into a digital, hence algorithmically manipulable, version.

This means that massive digital corpus creation, which is now slowly underway, is a prerequisite to bridge the methodological gap. This first step needs a strong boost. Only then quantitative methods can be fully exploited.

But we are not only dealing with quantitative methods. Incomparable to the standard natural sciences, the main characteristic of the humanities is the battle between competing interpretations of historical, moral, political or otherwise human activity. This means, for both quantitative as qualitative usage of the digital research corpus, that the digitalising process of sources in the humanities demands a clear insight in prospected usage. In other words, the structure of the digital corpus is of great importance. The methodological

challenges in structuring digital resources define common tasks in the humanities.

4.2- Quantitative studies as a rule?

In the humanities, certain schools try to be quantitative as far as possible. In the US, political science is completely quantitative. Also, in economy, psychology and sociology, important currents try to be as quantitative as possible. A large part of the sociological research is based on well-formed standardised interview protocols that can be analysed by the omnipresent Statistical Package for the Social Sciences SPSS [16]. In those fields, the corpus of research data is well-defined and made fit for formal logical operations. The enthusiastic embracement of sociologists of the UK E-science programme stresses the potentials for quantitative research as can be read in the large Economic & Social Research Council (ESRC) study by Cole et. al. “*Grid enabling quantitative social science datasets- A scoping study*” {Cole03}. Unfortunately, in this elaborate and interesting report, no figures are given about the needed bitstreams and storage capacities. The issues dealt with are very important for all types of collaborative work, but the specific demands for social sciences are not made very clear. In another enthusiastic, consultative, study by Fielding {Fielding03} the claim is made that “*it is timely to assess the potential returns to qualitative methodology and research from GRID and HPC technologies*”. It is left to the reader of the present report to compare the issues discussed below with the idea that a return to qualitative research is waiting for Grid computing. In a third ESRC report on funding, Woolgar {Woolgar03} defines e-science as “*science increasingly done through distributed global collaborations enabled by the Internet, using very large data collections, terascale computing resources and high performance visualisation*”. Typically, all ESRC reports were written with questionnaires and interviews as experimental data, data that can be dealt with statistically.

In the humanities in general, we just started with the creation of digital corpora and it is still an open challenge how ICT will assist in new quantitative and qualitative methods. An answer to the call for distributed Grid computing and HPC technologies can only be fully answered when a large number of substantial corpora are electronically available. At present, many of the problems of digitalising the research corpus in the humanities are not yet solved. Therefore it goes without saying that this induces a feeling of inappropriateness of the ICT push. After the discussion on these conditional issues, this report will conclude with suggestions that might prove the appropriateness of ICT and the possibilities for E-humanities.

Given the limited time in which this report had to be written, a deliberate choice is made to concentrate on those disciplines in the humanities that do not (yet?) have datasets that in structure and size mimic those of the natural sciences.

5- Five levels of processing

Below, we discuss the research process instead of providing an overview of the different disciplines. This way we can better understand the common and specific issues, in order to arrive at a better understanding of the “*methodological commons*” that may lead to a programmatic approach of initiating, stimulating and facilitating E-humanities.

In general we can distinguish five different levels of tasks.

5.1- Defining the information object: the naming of (experimental) data, artifacts, or sources

This phase is characterised by discipline-dependent protocols. An archaeological excavation demands different tools and methods than making an inventory of a mediaeval cloister library. The most obvious reason being that an excavation can only be done once, which makes it the antagonism of Popperian science as falsification is impossible by definition. In all disciplines we have descriptive methods identifying (naming) the various objects, including their physical status (pottery, census data, speech,...) and describing the context of their provenance, original and present location, etc.

To a large extent, we deal here with building resources. Such a resource consists of the original artifacts (texts, clay tablets, paintings, utterances, etc.) and their electronic representations and descriptions. This can be compared to a collection of stones and the database of related crystallographic data. Obviously, in both cases the real thing is located in one place, and the digital representation is in principle world-wide available. The difference is that crystallographic methods pertain to all solids, whereas measurement systems in the humanities are much less general. For that reason, a great effort is presently underway to describe and represent, in some electronic form, the current holdings of corpora and collections. In the strict sense, this is E-based humanities and not yet E-humanities.

5.1.1- Naming entities

It is an old cognitive truism that by naming something, the human being is able to think about it and to manipulate that something. The real problem starts from here, as the objectification of that something turns out to be highly “theory laden” as well as cultural- and language-dependent. The role of science is to disambiguate knowledge in such a way that we find a common ground for further enquiry. A good illustration of this problem is already given in the different names of this process of translating notions and concepts, to form a common understanding. The English word *translation* suggests a linear shift along an axis, while the Dutch *vertalen* indicates that something is expressed in language and the German *übersetzen* tries to describe a process of transformation. In language studies, the difference is exemplified by the analytical approaches of the linguists versus the more cognitive cultural approach of literary studies. An actual example is the new Dutch oecumenic Bible translation, where the discussion is between the Reformed Church adherents who want to stay close to the concrete meaning (since Luther, the truth is in the text) and the Catholics who emphasise the pious experience of reading the Holy Scripture {*Hij slaakte een kreet en gaf de geest* (C) versus *Hij slaakte een kreet en blies de laatste adem uit* (R)} (Netwerk04).

This example already indicates the inherent problem in language studies: to what extent can we describe a language including its phonological and physical/ physiological expressions and to what extent can we identify meaning, which can only be done in a comparative way in a certain socio-historical context. Language studies of endangered languages (see also 5.2.4) therefore immensely gain by using multimedia recording equipment, in contrast to the study of ancient languages which must restrict itself to analytical approaches (Talstra92). In both cases, ICT turns out a great help and stimulus.

In a digital representation we have to name the object as precisely as possible, taking the right granularity into account. A painting is an object, but also a whole document or just a morpheme. Only then we can manipulate the digital object, as the digital object is not more than a linear bit stream, containing descriptive information. This is distinctly different from dealing with an analog object. A serious problem then is the fact that there exists a clear challenge to be investigated: to what extent is naming in a taxonomy and thence manipulating data of any kind sufficient to induce novel research. The present stage of ICT is still very much based on a semantic approach of entity definition with its attributes, and rather less on the issue of structural analysis and processes, that are not necessarily hierarchical. It has to be stipulated that this is an intrinsic problem, as is well illustrated by the whole discussion on artificial intelligence, expert systems, semantic networks and, today the semantic web, where the classification of nouns (objects) is much further advanced than the understanding of classes of verbs (processes) (Fellbaum98, 02). The shifting context of a source, text, or image cannot yet be properly defined. The whole issue of so-called ontologies (thesauri with rules) is therefore essential (see 5.1.3).

5.1.2- Descriptive languages

The descriptive level of information gathering is primarily semantic and is well suited for so-called descriptive languages, a field that took off in 1986 when the Standard Generalized Markup Language (SGML) became an ISO standard. In this standard, a strict split was made between the markup tags that define content and the ultimate layout and representation into a media of choice. From a socio-historical point of view, it is interesting to realise that it was the first watered-down application of Hypertext Markup Language (HTML) that boosted the World-Wide Web (W3) initiative in the early nineties, which success forced developers to go back to SGML and design a subset that is optimized for the Web environment, named the Extensible Markup language (XML). It is this large international endeavour, coordinated by the W3 Consortium (W3C) [2], that enables initiatives in integrated methods for naming and therefore manipulating objects. The dynamic environment of the W3 world creates a fantastic boost to the management of existing and new to-be-discovered information objects. In the first place, it enables the naming (tagging) of entities independently of their lengths/size or physical appearance. The same object can be endowed with many different “tags” (e.g., language, corpus name, functional name such as paragraph, etc.), that might be used or stripped away, depending on the usage. Though, Information Retrieval (IR) research based on XML-tagged corpora is still in its infancy [17], it is obvious that when indeed large corpora become available, IR techniques will readily make use of the segregative power of XML-coded material. Some so-called search engines already use the limited structure of HTML-coded text. Independently of the W3C, a real community endeavour emerged in 1987: the Text Encoding Initiative (TEI). The TEI [18] is an international and interdisciplinary, SGML and XML compliant standard, to “*represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent*”.

A caveat with XML is that only nested tagging is possible (in mathematical terms it is a acyclic graph: a tree structure), it does not allow for overlapping tagging. The last technique is essential in deep markup in, e.g., literary texts when in a text, distinct segregations are useful for the establishment of more views of the same source. This issue is dealt with in 5.4.1.

However, semantics is not enough: as in chemistry, we can break up a molecule in its constituting atoms. The real thing can only be retrieved by introducing various kinds of chemical binding. In the W3 environment, binding is called hyperlinking. The present-day situation is that hyperlinks are still one of a kind. They point to somewhere without informing the user why. This indiscriminate way of binding different objects blurs the real relationships. The research on named hyperlinks within a more-or-less formal way, e.g., by using argumentational or structural schemas is also in its infancy (Kircz00). This issue is further discussed in relation to authoring tools in 5.4.2.

5.1.3- Indices and Ontologies

The age-old discussion on whether an object is endowed with attributes or whether the object is defined by the union of a certain number of attributes, is receiving new impetus in the age of digital storage. Automated systems are well-suited for administrative tasks. For that reason, a great number of index systems are now under construction. Two aspects have to be addressed by designing such systems of information about information, normally called metadata. In the first place, we have to make a distinction between nominal metadata, which can be standardised for large collections of diverse objects, and domain specific data that only pertain to a well-defined context. In the second place, we have to obtain an understanding about the various relationships of these metadata.

With regard to the first point, obviously the metadata content will be different in different fields, but the way the metadata is wrapped, the pointers to the metadata record itself, the pointers to the digital objects associated with the information object should not be discipline-dependent but is rather part of the e-humanities basic infrastructure.

In many systems, the two kinds of metadata are mixed and domain dependent metadata (keywords) are part of a larger metadata model. Given the great advantage that XML coding provides, a true explosion of metadata schemes and therefore proposals for mapping such diverse schemes, can be visaged in all disciplines, in parallel to an attempt to define common umbrella notions as is the goal of the so-called Dublin Core (DC) [3], and schemes that try to comply with the DC approach such as those in the Open Archive Initiative (OAI) community [19]. The important issue here is to what extent is the metadata scheme a general bottom-up scheme, or to what extent you deal with a collection of specific vocabularies. An example is the field of language documentation where, e.g., the group of sign languages comprises a different tradition in description than spoken languages.

In particular, in the fields of collections and archives, we see the emergence of very elaborate schemes. Often they integrate a large number distinct types of information. An example is the Dutch-based system for Art History, Iconclass[1]: *“a collection of ready-made definitions of objects, persons, events, situations and abstract ideas that can be the subject of an image. Iconclass organizes iconography into 10 'main divisions', each containing hierarchically ordered definitions”*. A recent, highly modular and flexible, example is Sepiades (Klijn03), a system for cataloguing photographic collections that is partly based on the General International Standard Archival Description (ISAD(G)). An overview of metadata standards for cultural organisations can be found on the website of Digitaal Erfgoed Nederland [4]. Also here we have to make a distinction between metadata skeletons, which we would like to standardise as far as possible and the meat on the bones that is species specific.

From a semiotic point of view all those denotating schemas are constantly in flux as soon as the context of use changes. The (sub)domain-dependent connotation is then again formalised in a new scheme. Here we are back at the early works of the semiologist Roland Barthes (Barthes64), that every denotation of information is in a context and that if we systematise the meta level of the context, we create a new denotation system that again is context-dependent and so on. Unfortunately, again and again initiatives for world index systems emerge.

Within the XML world, this issue is now addressed by the so-called Semantic Web [11] initiative. This new field is enthusiastically embraced by all disciplines that deal with large data-sets or holdings in almost all fields. It defines itself as: *“The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming. As was the case in the old days of library thesauri every discipline is struggling with its own specific keyword systems”*.

Therefore, the second issue mentioned above: the relationship between the various metadata becomes a crucial factor. Standardisation rules and transparent grammars are essential ingredients for the creation and exploitation of digital repositories. These kinds of taxonomies are nowadays called Ontologies, systems that formally model a knowledge domain (see (Vickery97) and (Gilchrist02) for the history and clarification of this term).

In technical terms, these issues are dealt with in so-called Web Ontology languages, of which OWL is one of the latest standards, that goes beyond the simple tree structure of RDF (Antoniou04). *“OWL is used to publish and share sets of terms called ontologies, supporting advanced Web search, software agents and knowledge management. OWL - the Web Ontology Language - provides a language for defining structured, Web-based ontologies”*. The ontology development exactly reflects the old issue of the tension between formal schemes that can be filled with domain-dependent content and the level of unique domain-dependent relationships between entities and attributes as well as between entities as such. This field of research is of the greatest importance, as all digital sources need proper indexing before they can be fully exploited. It is also here that the step will be made between E-based humanities, the electronic availability of index systems of all kinds, and E-humanities where we will apply computational methods for creating and using well-defined objects and relationships.

In the digital library world various collaborative initiatives are in full development such as the Metadata Encoding and Transmission Standard (METS) [39] and advanced schemes as MPEG-21 Digital Item Declaration Language (DIDL) (bekaert03). These schemes deal with XML-based data structures that not only deal with the item itself, but also with intellectual property management, rights management and the issue of transparent exchange and content packaging.

We can conclude this section with the firm conviction that the field of metadata schemes, descriptive and ontology languages is crucial for the creation of full-blooded digital resources.

5.2- Writing, rewriting and presenting

Knowledge obtained from a source, be it an archive, book, or experiment remains a dead body of information until a scholar uses it for a scientific report of any kind. This process of creating knowledge can be a speech act, or a symbolic representation. In distinction to the past 400 years of scientific scholarly communications in writing, the electronic revolution enables all kinds of representation for conveying information and knowledge.

It is immediately clear that comparative studies now gain a great deal. Concordance studies between texts become much easier if the source texts are digitally available. Images of manuscripts and their wordprocessor transcriptions in both the old spelling, and, if wanted in modern language, can be easily displayed in one view. A great many cultural institutions and libraries are in a process of digitising their holdings in a standardized way and making them available for amateurs and scholars.

The huge endeavours vis-à-vis the national heritage, dealt with in section 5.3, form the prime example. They set the stage for new ways of doing science.

The crucial novel aspect provided by ICT is that due to a fine-grained digitalising mock-up analog information can be presented next to symbolic information. In print, we were faced with symbolic information in the form of written text, augmented by analog information in the form of images, whilst sound only became an auxiliary aid with the inventions of sound recording equipment. In the digital world, images (still and moving), as well as sound can be recorded with a resolution far greater than the physiological threshold. This means that in the reporting, the domination of text is under attack. In the biomedical world, we already see the upsurge of imaging techniques as primary information sources. Here, just as in art history and archeology, the image has the primacy and the text becomes the explication of the image, in contrast to traditional abstract science where the image is called illustration or illumination. This means that the mutual interaction of sound, image, text and soon simulations together, create the scientific argument, and therewith a complete new scientific rhetoric.

We can distinguish the following distinct processes:

5.2.1- Rewriting of existing and new textual material

Here we deal with existing material that has to be translated into a new form. In its most simple way this is done by Optical Character Recognition (OCR) techniques, which work well for well-printed material. More complicated is the study of how to digitise handwritten text. Within the NWO Catch programme {Catch04}, a research proposal is incorporated. In both cases, we are representing existing textual information into the new medium in order to facilitate automatic procedures that help in, e.g., concordance studies and enable IR techniques. Though of great importance in itself, it is in fact only a preparatory phase towards complicated computer-assisted text analysis of manuscripts. This important work belongs to the phase of bringing all existing knowledge onto a common digital platform, thereby enabling novel types of investigations.

Furthermore, the enrichment of existing text corpora with XML coding belongs to this important preparatory phase. A good industrial example is the incorporation of back volumes

in HTML (partly XML) form into the scientific publisher's holdings. Elsevier's Science Direct [33] is presently the most advanced example.

The issue of preparing new scientific textual material directly into electronic form is an important new issue, as in these cases we already make use of predefined electronic environments, for instance in the form of wordprocessors. It goes without saying that a deep analysis of the structure and presentation of already existing material will be an important ingredient toward the definition of new standards for works that are still to be written. However, there is no reason to believe that textual information in electronic form will be of the same form and style as textual information in printed form. For the case of scholarly publications, see Kircz (Kircz01, 02). This brings us to the pressing issue of authoring systems that will be elaborated in section 5.4.2.

5.2.2- Image digitisation

On the level of images, we are mostly still in the early phase of digitising existing collections using an elaborate descriptive metadata system. In doing this, we encounter various caveats.

A- Whilst the analog originals remain the same objects, as long as they do not erode, the digital representation can keep its integrity forever, provided the binary code can be read and interpreted. This problem of digital preservation falls beyond the scope of this report but is a main societal concern for civil administrations [22], the national and local archives, the cultural heritage institutions, the special collections of the universities and their museums, and the Koninklijke Bibliotheek, National Library of the Netherlands (KB) [23].

B- The changing standards and techniques force collection holders to think twice before they start major operation. At present, the Rijksmuseum [5] is engaged in a huge project to create a fairly complete digital repository of its holding. The guiding policy is one of best practise, as today, standards are in a state of flux. Will the open source png standard ever replace the Tiff standard? What will be the standard type and size of representation in 2008 when the renewed Rijksmuseum will be open again and only 8% of its holdings are on display? Even if the mission of a museum is limited to the primary goal of showing its holdings and allowing the general public and scholars to inspect the originals on request, the digital representation will remain in flux.

A typical pragmatic decision is that all objects will be available in 2D form, including the large holdings of 3D objects. Here and at the current project for the Allard Pierson Museum [6] high resolution Tiff files, including a colour management system to guarantee quality, is the best practise. If necessary, more images of the same object are taken. 3D imaging is still in its infancy, so it is known that in some years time, part of the collection might be re-recorded again, then in 3D format. This will become the new, normal practise for Museums. A notable exception is the Jewish museum in Vienna [7], where holographic images are on display. Though very impressive in their size, the question remains in what setting this kind of representation enables a different and additional understanding of the message.

In the field of film and video recording the work on descriptive languages is now a very active field, though still mainly the work of computer scientists.

5.2.3- 3D visualisation

It goes without saying that 3D imaging and holography will become an important representation method of 3D objects. In archeology, 3D imaging is already a standard (though not fully integrated) practice. The Internet Journal: Internet Archeology [8] is a good example, where also Graphical Information Systems (GIS) are extensively used as a research tool. It is also telling that unfortunately, as in many other e-journals, the cultural threshold to publish in such advanced, high quality journal is very high. Clearly, only the most enthusiastic and advanced groups populate the new pastures. As mentioned earlier, in these examples of E-humanities, the lack of education and familiarity pays its toll. In the GIS field, the Virtual Reality Modelling Language [10] is an important tool. Cross-fertilisation between the Humanities and GIS intensive disciplines like geography and geology and parts of chemistry where 3D molecular modelling is daily practise, can certainly be enhanced and encouraged strongly. The use of GIS for historical information science is also one of the key issues in the review by Boonstra et.al. (Boonstra04).

On another example is Cave technology, a new computer-based way of dynamic 3D visualisation providing true stereoscopic imagery [15]. It started enthusiastically as tool for industrial design and the understanding of complex 3D structures, but also here, the technological success could not be easily mapped onto to the demands of people struggling with visualisation problems. Again, the mismatch between technology driven inventions and the explication of intrinsic demands of representing information and knowledge, becomes clear.

5.2.4- Sound and film

The registration of music is of course taken the lead in the development of new techniques. Also here, it is most important to know what standard is necessary and sufficient for research purposes. A “lossy” compression technique as MP3 is fine for recreational or many educational purposes, but fails if faithful analyses are wanted. We may note that in the field of the spoken word, the groups working on endangered languages pay great attention on the recording of sound in combination with film or video. Digital Audiotape (DAT) recording is the rule here. In the language documentation world it are not only the sound files and their phonological transcriptions that count, but also field notes, video and film recordings of gesture and the interplay between the story teller and his/her audience, as well as a whole collection of legacy data that might go back for centuries. In the international, arena we have various large collaborations such as the Electronic Metadata structure for Endangered Languages Data (EMELD) [24] that tries to promote a consensus between linguists about naming, structure and content. In the ambitious Rosetta project [25], that tries to build a “near permanent archive of 1,000 languages”, audio files are an important ingredient. Also at the Meertens Institute for Dutch language and Culture [31] we deal with large quantities of audio recordings of Dutch dialects, many of which are still on old to very old sound recording media. Digitisation is then a result of the goals of a research programme and not a goal in itself.

In language annotation, we deal with different data-structures since we have a time-line for the audio file, and parallel to that orthography, the lemma, the part of speech, the phonological transcription, etc. This means that we deal not with a tree structure but with an

elaborate lattice. A good example how this complicated issue can be tackled by using XML stylesheet techniques, XSLT [26], that allow for multiple disjoint representations is given at the EGA WEB ARCHIVE: “*An endangered languages documentation initiative*” [27]. The Spoken Dutch Corpus (Corpus Gesproken Nederlands) [34] is a large database of contemporary Dutch with an elaborate tagging that can be browsed with a special tool, Corex, that allows for queries for tokens as well as statistical queries in the annotation data. Although film and video registration is an important tool for language studies, as well, as anthropology, ethnography and behavioural science, in the course of writing this report no specific usage could be identified that is unique for the humanities.

5.2.5- Thinking

An even more abstract level of knowledge representation can be seen in parts of philosophy. Here the reasoning is represented in formal computer code to simulate and explore thinking processes. With this kind of experimental reasoning research, the consequences of formal decision processes can be analysed and tuned. In the work of the Canadian W. Maurice Young Centre for Applied Ethics [9] (director Peter Danielson), rule-based simulations are tested. The ground work of these types of programmes can be found in statistical mechanics and game theory, but the rules of the “software agents” do not have to adhere to the laws investigated in the natural sciences. A Dutch example is, e.g., the PhD thesis of Hans van Ditmarsch (Ditmarsch00) on the board game Cluedo. Reasoning patterns are now represented in software. A long way from the Athene marketplace. AGORA, a web-based framework for ethics education targeted at engineering students, is supported by SURF [28]

This type of modelling proves that computer programs are becoming a mature representation of reasoning, though still formal and abstract. Computer programs are therefore evolving from calculation aids to serious knowledge representations in themselves, that need the same attention and care as other knowledge repositories.

5.3- Storing information

5.3.1- The digital archive

Eamon in his “*Science and the Secrets of Nature*” (Eamon94) explores the thesis that real science only became possible after all existing information, even irrelevant or plain wrong, had been printed, because only at that stage did it become possible to compare and sift through the various data, recipes and descriptions. This transposition of all available information from old sources into a new, unifying one, provides a good metaphor for what we currently see happening with the Internet. As before, the medium does not pre-filter the information: astrology and pornography are as readily available as cosmology, or the classics of world literature. Yet, the very process by which virtually all human information becomes available in a standardised form, facilitates sifting and comparison and thus may lead to new advances in all areas of academic endeavour.

At present, we are in the first stage of making existing analog collections available in digital form. This is a society-broad, but nevertheless highly uneven phenomenon. At an advanced stage we encounter the scholarly journal literature, though only a limited number of

publishers do indeed produce their material in fully encoded XML form. On the same level, we see highly indexed and well-structured research corpora that allow for quantitative research as well as qualitative analyses. A good example is the recently published *Biblia Sacra* [12], an electronic bibliography, allowing its users to request information (text and reproductions) on bibles printed in the Netherlands and Belgium. It is interesting to compare this with the fascinating site *De Bijbel in de Nederlandse Cultuur* (The Bible in Dutch Culture) [13], which is an encyclopaedic work with an incredible amount of textual, pictorial and even audio information. In a way, the last example is a proof of quality how a single digital platform enables the various cultural expressions on a certain subject. The first example, is in fact a step further, as its deep systematic metadata structure enables a high level of manipulation by the user. Here a real research corpus is created. In the digital production centres of some of the university libraries (e.g., Amsterdam, California Digital Library, etc.) and the KB serious programmes are in place to transpose their special collections to an electronic research corpus.

In the public domain, the government is a great promoter of an electronic infrastructure (see for the relevant documents the Government portal {Keo04}), but this policy is not yet grounded on the construction of a fully digital interoperable environment. In particular in the national heritage world we see an emphasis on presentation per se and not on structured information. A PWC Consulting report {PWC03}, made for the department of Education, Culture and Science, rates the ICT usage in Museums as high. But the real issue is not the amount of equipment, but the level of internalisation of the methodology inherent to an e-based environment. The great variety of quality and coherence between the corpora of museums can be quickly experience by visiting the website of the *Cultuurwijzer* [14]. Or, as one of the interviewees commented “*up to now the many websites of museums are built with the showcase as metaphor*” It is fully driven from the analog corpora and not based on any possible use by amateur or scholar. Inconsistency and incoherence, even within an institutional digital corpus, are almost the rule. This orientation on web presentation instead of a sound digital production of object representations, of which the web site is only a rendering, is also the cause that in an E-learning environment no direct hyperlinks to an object, that is displayed in a website, is feasible.

5.3.2- Dedicated storage environments

As implicitly mentioned already above, we have encountered a clear methodological choice to what extent a digital resource is created for the general public (including researchers) and to what extent is it a primary research tool. The enormous *Perseus Digital Library* [21] started as a repository on the Antic world is explicitly designed to attract the general public (Crane98). The same philosophy can be found in the recent Dutch report on the public function of archives {Dijken03}. In general one can state that one of the intrinsic characteristics of a digital archive is that it must be approachable by scholar and layman alike. This also holds for sound and film archives of cultural, literary, musicological, linguistic value. Digital resources are by definition available to everybody. Hence those sources that might be interesting for a larger audience than the initiated must at least have a layer that is understandable and usable for the interested lay person. Obviously, this also demands a further step in streaming audio and video services within the SURF-net environment.

This brings us to the discussion on dedicated data structuring and data base management systems (DBMS) that are source compliant. In particular, in history the discussion on dedicated data storage systems has already been going on for decades. Historical sources are a prime example of heterogenous data collections that are research corpora without a fixed structure. A key characteristic is even that sources are constantly reinterpreted in the light of changing cultural or scientific approaches. Historical data have an inherent fuzziness and interpretation is part of the craft. The intrinsic context sensitivity of the historical databases (Thaller89) is in a permanent tension with standard office automation solutions. A massive overview of all aspects of historical information science is given by McCrank (McCrank02). Boonstra et. al state that *“Historical data is to be administered as pieces of text, without any assumption about its meaning. Meaning depends on interpretation, which is a fruit of historical research. Therefore, data should be entered in a source-oriented way (keeping together in a single file what appears in a single source document), rather than in a program-oriented way”* and *“The typical historical database management system would be a hybrid between a traditionally structured DBMS, a full-text retrieval system and a document retrieval system”* (Boonstra04, p44). The same authors explicate the difference between literary research and history as they state that *“Literary research has been focussed on critical editing and analysis of the text themselves.....in contrast...historical research problems pertain less to the texts themselves, but are more related to the historical reality beyond the documents handed down”* (ibid. p51).

In the present discussion on a possible national data archive {Dijk04} and various important initiatives to link a great many sources into an integrated repository such as the Alf@net project {alf03}, it goes without saying that the structuring of the various repositories and data banks demands serious attention. The fundamental issue is again, do we deal with a general repository, bringing many different fields together, or do we deal with a research corpus, where the tools are discipline dependent. In a (national) data archive an archeologist will require a very different set of tools than a classicist. This issue is intertwined with the mentioned challenge of the balance between metadata schemes or skeletons and database management systems on the one side and the content of the (metadata) fields on the other side.

5.3.3- Chaining

A related issue that needs serious attention is the analysis of the logistic process of information usages in heterogeneous environments. The tardy development of a system of medical patient dossiers is a good example. We deal with textual and non-textual information that changes over time and where different data have different levels of confidentiality and integrity demands. Though not yet incorporated in the E-science discussion, the digitisation of lawsuits is now making its first steps. Unfortunately, the Dutch courts now mainly dump textual material in scanned form onto a CD-rom, which is of little help. However, some interesting projects on chaining digital information are now in place, such as experiments with the forwarding of XML-coded police reports. Aside from the technological aspects, a major issue is, just as in the medical case, the great variety of kinds of information, their confidentiality, integrity demands, and indexing, which put special requirements on storing, access, updating and retrieval. These two cases are extremes in their complexity, but are excellent examples of the intricate constraints in fields like, e.g., civil law, toward a future

electronic environment. In contrast to the blossoming field of intellectual property right studies in cyberspace, which is describing societal aspects of E-science, here we deal with the transformation of the judicial world itself due to electronic means, a field that has hardly started.

5.4- The analysis of digital corpora

In the previous sections of this chapter, already many issues are dealt with that prepare the next stage in humanities research. Naming the objects of investigation, structuring them and loading them in a common memory are all prerequisites for the emergence of E-humanities. We named this preparatory stage, E-based humanities. As elaborated on at the beginning of this report, E-humanities means that the whole corpus under investigation has a structured digital representation. This digital version will entail, as far as possible, all types of digitised information. Structured text will exist next to a high resolution image of the original handwriting. Phonological transcriptions exist next to the sound files and the orthography. However, in many cases, the analog artifact will still be available for inspection somewhere. The essence of E-science and E-humanities is that this somewhere can be far away and even unreachable for the investigator. Nevertheless, the electronic representation enables advanced research. In this report only a few examples will be given, as this 4th task in the list pertains to domain specific research and only becomes a typical SURF issue if SURF is able to assist and to facilitate it. In the conclusions, we will come back to this point.

5.4.1- Two poles

Right from the beginning of our discussion in possible E-humanities, we have to make a distinction between those investigations that use humanities corpora for the development of methods and techniques and those investigations that have the understanding of the corpora as such as its primary goal. The dialectic between these two poles is important to keep in mind, in particular as the first pole is to a large extent technology driven and is often populated with people with a computer science background, whilst the other pole is populated by mainly traditionally educated scholars.

A good example is the community of language and speech technologists collaborating in The European Network in Human Language Technologies (ELSNET) [29]. In this research the aim is to define common technological denominators that enable to analyse and use language as a communication tool. Hence, the industrial interest in this field is also high. The goal is not a full understanding, but to reach a sufficient understanding that meets the requirements of a well-defined problem. Speech recognition and speech simulation, e.g., for railway or telephone services are a good example. Also computer-assisted translation technology is part of this pole.

In the same league, we can place the Natural Language Processing (NLP), Information Retrieval (IR), many metadata and ontology architects, and image recognition researchers. Their main goal is the development of potent algorithms that can be applied to a great variety of “raw material”. Here, one is after universals and not necessarily after the particulars of a well-defined problem area. Statistical methods are therefore often important tools.

It goes without saying that the precise problem definition of the corpus experts is a needed ingredient for these technologists. Unfortunately, the disparateness between the poles is so large that often communicative problems prevent a constructive and smooth collaboration. So

often projects are more “*A gallery of possibilities*”, as one interlocutor named them, than a direct asset for the underlying field. This is clearly also reflected in the already mentioned discussion on historical information science.

On the other pole, we find beautiful examples of ICT use in analysing complicated issues. One example is the work of Willard McCarty (McCarty03b). The object of research is Ovid’s *Metamorphoses*, in which the literary phenomena of personification of non-human objects is the issue of research. The author identified relevant linguistic figures and encoded this information. He coins his 55,000 tags for 12,000 lines of Latin hexameter “thick” encoding, in contrast to “deep” encoding that tries to reveal an underlying grammar. With this declarative encoding an attempt is made to reveal, with the help of a standard relational database, the underlying structure and grammar. This type of encoding text is really intellectual labour and fits De Belder’s dictum “*I tag, therefore I think*” (Belder93). The work of Steele (Steele96) on Shakespeare derives from the same school as McCarty. Here we deal with the difference between the analyses of a play in distinct acts, scenes and lines versus the continuing dramatic action of one of the personalities. In all cases, we have various tagging schemes interwoven. Formal XML is not able to handle this, as well-formed XML is nested. Although not used by the authors mentioned, this example might be a good candidate for an XSLT approach. The same technique which we encountered with the Endangered language community in mapping various description schemes of the same utterance.

This is not the place to discuss this example in depth, but it is sufficient as proof that, in principle, good communication between different field can create *methodological commons*, where humanities and computer science can meet and cross fertilise each other.

A problem spanning all fields is the fact that, as was already mentioned in section 5.1.1, present day technology is mainly based on semantics. In the search for syntactic patterns, e.g., in dialectology, or modern youth language, standard commercial search engines are of little help. As in this type of research inflections, word order and articles are essential, while the search engine philosophy is built on stemming, and deleting articles. The dynamic research in current speech communication demands the ability to scrutinize the various genres of internet communication, such as web pages, news groups, chat boxes and audio/video files as they arise.

5.4.2- Multimedia tools

At the end of this chapter it is helpful to summarize some of the tools necessary for scholarly work in an ICT-based environment. As this report is not the place to list research projects only a scant listing will be given.

A- At the basic level, metadata schemes are needed to describe information. The development of standards and their related software is a permanent concern, as long as hard- and software environments have not yet reached long-term stability. In particular in the non-text area, this work is still in its infancy, which does not mean that in the text area, we are already beyond adolescence.

B- All work starts with the availability of a corpus. Hence the corpus has to be encoded in such a way that segregation of the richness of data is made easy. This induces the discussion

on the granularity of encoding. As mentioned in the case of repositories, not all users need deep or thick encoding. In technical terms, this can be expressed in the notion that the Document Type Definition (DTD) that defines the document elements does not have to be the same for the production stage and the presentation stage. For example, the most elaborate production DTD for scientific journal articles is presently that of Elsevier BV [32]. It is clear that not all fields are useful for searching or retrieving. It is still an open question to what extent a presentation of a single paper or a whole repository, such as Science Direct [33] in the Elsevier case, has to be endowed with the full production DTD. At present, Science Direct uses only a very small subset. The same discussion holds for all digital holdings, as not all information that can and will be “tagged” for completeness sake, because you never know what is needed in the future, is necessary in actual research. Search tools based on the underlying coding are still to be developed.

C- Creating new works means writing, which now entails more than text only. Only the author is able to correctly name the objects (s)he uses. Hence, already at the authoring stage information must be given the appropriate name, in order to avoid ambiguities. This means that authoring tools are indispensable. Not only the standard information such as the correctly spelled name and address, but also keywords and structuring information, such as named hyperlinks, can only be exactly defined by the author. Various experiments with authoring tools are underway, though no full-scale development targeted to a transparent mapping of the author’s file and the repository exist. An excellent recent comparative overview of ontology-bases authoring environments is given by Oliveira (Oliveira04), which clearly indicates that it has all just started.

D- In an integrated database, images, sound files, and text are all available. At present, images (and sound files) are retrieved by their descriptors (file headers, captions text, textual metadata). As image repositories become more and more important, also other IR methodologies will be needed. Pattern recognition research based on, e.g., colour, shape or texture are now under development. Their integration with traditional IR methods is certainly a requirement for the near future. A central theme is to understand the way they are successful and what their intrinsic weaknesses are. An good example is the use of arrows on biomedical images to indicate salient details. This arrows are most disturbing for any form of image retrieval based on pattern recognition techniques. By knowing the intrinsic strong and weak points of a certain technique better “instructions to authors” can be implemented that pre-empt known complications.

E- As already mentioned in 5.2.4. the usage of sound and video files is mounting. Stream techniques become important for browsing this information.

5.5- The reading and using

At the end of the day, as everything is said and done and more is said than done, we deal with the layman, student and scholar who is interested in the data and/or the results of the research. This aspect falls outside the scope of this report, but for completeness sake some observations are listed below.

A- As already mentioned in the first parts of this report, the differences between the natural sciences and humanities are reflected in the way the whole ICT infrastructure is set up. People who deliberately do choose for a non-technical discipline are put off if their dormant ICT needs are approached by jargon, straitjackets they don't understand, and claims that are not based on an intrinsic knowledge of the research quests.

E-learning therefore can play an incredible important role in integrating intellectual challenges with new technology. Putting two versions of a manuscript next to each other to do concordance studies is doable on a clean desk, with a pencil and paper. But concordance studies in an electronic environment are much easier if the methodological understanding of the why and how are in place. ICT then becomes the realm of scholars that know how to play in a research environment that is shaped according to their methods. Elaborate websites can also play an important role in the day-to-day education of students and citizens. A good example is the site of the Dutch-Belgian Taalunie [36], with its platform on language and Speech technology in which many organisations and companies collaborate [37]. In developing a *methodological common* to the humanities, the ICT tools have to be targeted to the issues at stake. We mentioned already the discussion on source oriented database management systems. After a clear structural analyses of the particulars of the various domains, we can design the common platform.

B- Reading hypertext in a multimedia environment is, like reading a traditional book, closely linked to authoring such a work. If people are confronted with clearly structured electronic files and at the same time have proper authoring tools, the acceptance and understanding of electronic-based science become obvious. Also as today's writing is the historical corpus of tomorrow, the authoring of new corpora in the humanities has to go according the latest standards of data management.

C- In, e.g. geology and weather forecasting a Collaboratory is defined as: "*The combination of technology, tools, and infrastructure that allow scientists to work with remote facilities and each other as if they were colocated*" (Lederberg and Uncaper, as quoted in (Moor99. p30)). This type of work will clearly emerge in the humanities as soon as a critical size of digital corpora will become available in structured form. International programmes will become standard and can learn a lot from the experiences of these trailblazers.

6- Conclusions

“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 its strategic plan ‘The heart of the matter’ 2003>6 we read; *“In ICT and Organization, SURF will focus its efforts mainly on the establishment of common database systems (middleware) for basic record keeping. The harmonization of systems in individual Professional and Academic Universities, and between these institutions, is a central issue in this”* {Surf03}.

These statements clearly define the SURF activities in the *methodological commons*. Within the Dutch landscape we see different partners that deal with E-science. NWO, KNAW and the universities deal mainly with pure research projects, while other, governmental, agencies try to enhance the co-called national knowledge-infrastructure. Digitisation and digital preservation projects are an important component thereof. An essential observation of this study is, that E-Humanities can only take off if digital research corpora exist in a sufficient number and in a sufficiently large size.

In this indicative 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. Due to time constraints, the emphasis was on the creation of digital corpora and qualitative research. These fields do not deal with huge data-streams, nor with the need for massive parallel calculations. However, this does not mean that these fields will not benefit from the logistical and computational infrastructure that will be built for the data deluge. Distributed digital archives and repositories, dedicated to E-humanities, will enable broader international collaborations and an increasing intensive data-traffic. But, one lesson is clear from the experiences so far, every community has its own grammar and vocabulary. Thus also in this context, before the capabilities of ICT seriously empower Humanities, the dialectic between technological possibilities and immanent research quests must be addressed.

Based on the overview presented, we arrive at the following ten points that emerged as preconditions for the blossoming of ICT tools in the humanities, in other words, the takeoff of real E-humanities.

Creating the preconditions for E-Humanities

0- As is made clear in this report, the creation of large digital research corpora is a prerequisite for all serious E-research. A strong national collaborative programme of universities, libraries and research organisations is needed. It is also clear, that such a digitalising programme is not a simple translation of old material into electronic forms. The iterative process of computer science research programmes and the formulation of the domain specific architectural demands in the humanities, calls for an integrated approach. The development of digital libraries including the identification, packaging and right management of digital content is closely related with such a digitising programme as many, if not all, digital objects must be retrievable via digital libraries. SURF can play a coordinating role.

On the level of education and instruction

These issues overlap with the SURF ICT and Education programme.

1- In the curriculum of the humanities the emphasis must switch from training in (commercial) office automation packages to a more methodological coupling of intrinsic research interests to modelling and architectural demands. Primary is the question, what do we want to know and how can we shape research. This important field is open for educational software environments and training in methodological thinking. The use of so-called Computer-Supported Argument Visualisation (CSAV) and other visualisation techniques will become an important asset. All type of humanities studies are in great need of a repository of well-structured and documented, textual and non-textual, data and tools to deal with the information in those databases. The university libraries, which already play an increasingly important role in digital learning centres, might become the coordinators.

2- Helpdesks are presently oriented toward the administrative and logistic infrastructure of the institutions, whose main task is the integrity of the network. Therefore in a research environment they are often considered as prison wardens who defend officially supported software, rather than colleagues who can explain why one software package is better suited than another for the problem at issue. It is not clear to many why some software packages (or office suits) are supported and others not. The Surfspot website [38] for ordering software does not provide any information about the pros and cons of a particular package. A real product analysis, based on which a choice can be made, before you find out that the package is not doing what you want, is badly needed. A clearing house function was mentioned at various time points. The best solution is not an “expert centre” with human experts who know better than the visiting scholar, but a system of well-defined product descriptions, their strengths and weaknesses and their documentation.

3- Information science can be considered as the go-between informatics and a particular field. In the Netherlands, only at Utrecht University, does Information Science go beyond text and library-based research. The traditional Computer and Humanities departments in the humanities faculties themselves are small and scattered. A new impetus has to be given by integrated projects based on the present-day trend to fund large-scale digitisation projects.

On the level of archives and repositories

4- As often there is no prior knowledge and training on how soft- and hardware requirements have to be defined, many an application makes a false start. The DARE-repository experience shows that collaborative activities will gravitate towards a working network that is supported by the various institutions and also used. In many research projects, data-sets are created or enriched. Normally they are not published with the textual reporting. As an extension of the DARE repository programme, a repository of research data would be a natural and most valuable extension. Such an initiative dovetails with the proposal for a National Data Archive for the Humanities {Dijk03} to be managed by the KNAW. It is an obvious conclusion that the complicated and heterogenous structure of such an archive is an excellent opportunity to start a DARE-like process for data repositories. As also in the other sciences data-repositories of all kinds are in the process of creation, umbrella architectural structures based on agreed, but (sub)discipline-dependent, standards need a closer look. The

methodological commonalities then extend over all disciplines and can co-define standards. The research into such standardised environments is a typical research project in itself that might fall under a NWO programme. This way, fundamental research in creating an E-based science will enable E-science. Among other things, this means that in the construction of digital libraries, archives and repositories a closer look at emerging standards such as the XML family and the source compliancy of the database has to be taken into account. This can become an elaborate project in its own right, where SURF, NWO, KNAW, university groups and the university libraries can collaborate.

5- The clearing house and help function for commercial software was already mentioned under point 2. However a lot of software is written or adapted for special purposes. Often after the closure of a project, this special purpose software is shelved and, even worse, in many cases without proper description. This waste of human and financial capital demands a closer look. An important point here is the assistance in “porting” well-working programs from their present environment, e.g., DOS, to a modern operating system. The creation of a well-documented program library, following the example of Sourceforge [30] “*world's largest Open Source software development website, with the largest repository of Open Source code and applications available on the Internet*”, is necessary. Quality certification based on past performance, with proper referencing has to be part of such an endeavour. Apart from the current DARE document-repository programme and the in point 4 mentioned data repository, there is an absolute need for a software repository. It is worthwhile mentioning that these kinds of repositories or libraries have already existed for many years in various fields. A good example is the Belfast-based Computer Physics Communications Program Library, established in 1969, that contains more than 2000 computer programs in computational physics and chemistry linked to a peer reviewed journal of the same name, in which the programs are extensively described [35].

On the level of dedicated software

6- Intertwined with the previous point, web services that enable the interoperability, to the extent that the data structuring allows it, can be set up, provided the underlying data repositories are built in such a way that they allow for web services. SURF can play a role in formulating the design and usage requirements. For the humanities, in particular the Web Services Description Language (WSDL) level, that describes the interface, and which is in cases of heterogeneous databases not trivial and the Universal Discovery and Description and Integration (UDDI) level, that takes care for the registration and search capabilities, will have their specific challenges.

7- Metadata structuring will remain a crucial theme for the coming years. Internationally many collaborative endeavours already exist that try to arrive at schemas for descriptive, administrative, and structural metadata, in particular in the field of Digital Libraries. It is important to team-up with these initiatives as E-science is international per definition, and hence demands common international standards and procedures. A closer connection between ontology (language) designers in all fields, linguists and philosophers have to be promoted. In particular, as first order logic modelling is not enough to cater for complicated ontologies, pragmatic argumentational approaches as well as more complicated modal logic schemes need attention.

On the level of user interfaces

8- The development of ontology-based authoring tools and collaborative software is an essential ingredient for E-humanities. Special care has to be taken that the metadata structure of these tools is in concordance with the structure of the repositories and archives that contain the research data and reporting. The same requirement holds for the related e-learning standards.

9- On the level of human-computer interfaces and the visualisation of information the development can go hand-in-hand with the natural and computer sciences. The development in computer-supported argument visualisation is an example. The usage of graphical presentations of economical, historical or census data, another. It is also worthwhile noting that the natural and computer sciences can benefit strongly from recent advances in language and speech technology, a discipline that, for a large part, traditionally belongs to the humanities in The Netherlands.

Acknowledgement

I would like to thank all interlocutors for their enthusiastic collaboration and discussions. In the report I have tried to explicate the issues mentioned as faithful as possible. I apologise for too harsh generalisations or unclear phrasing. In particular I would like to thank Leo Waaijers SURF-DARE project leader, who commissioned this report, for the stimulating discussions. I wholeheartedly thank Kurt de Belder, Jörgen van den Berg, Gert Goris and Steven Krauwer for reading the draft and suggesting improvements, and my friend and former colleague Keith Jones for scrutinising the English. All typos are mine.

Glossary

CDROM	Compact Disk Read Only Memory
DARE	Digital Academic Repositories
DAT	Digital Audiotape
DBMS	Database Management System
CERN	Conseil Européen pour la Recherche Nucléaire. Now: European Organisation for Nuclear Research
DC	Dublin Core
DTD	Document Type Definition
ESRC	Economic & Social Research Council UK
GIS	Graphics Information System
HPC	High Performance Computing
HTML	Hypertext Markup Language
ICT	Information and Communication Technology
ISO	International Organization for Standardization
IR	Information Retrieval
KB	Koninklijke Bibliotheek, National Library of the Netherlands
KNAW	Koninklijke Nederlandse Akademie van Wetenschappen (Royal Netherlands Academy of Arts and Sciences)
LHC	Large Hadron Collider
METS	Metadata Encoding and Transmission Standard
MP3	Moving Picture Experts Group Layer-3 Audio
NWO	Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research)
OAI	Open Archive Initiative
OCR	Optical Character Recognition
OSI	Open System Interconnection
OWL	Web Ontology Language
RDF	Resource Descriptor Framework
SGML	Standard Generalised Markup Language
SURF	Stichting Universitaire Rekenfaciliteiten
TEI	Text Encoding initiative
UDDI	Universal Discovery and Description and Integration
URI	Uniform resource Identifier
URL	Uniform Resource Locator
XML	Extensible Markup language
XSLT	Extensible Stylesheet Language Transformations
W3	World Wide Web
W3C	World Wide Web Consortium
WSDL	Web Services Description Language

References used (but not necessarily explicitly quoted in the text!).

R1-(Scholarly publication)

- Antoniou04 Grigoris Antoniou and Frank van Harmelen. A Semantic Web. MIT Press, 2004.
- Barthes64 Roland Barthes. Inleiding in de semiologie. Dutch translation. Meulenhoff 1970
- Bekaert03 Jeroen Bekaert, Patrick Hochstenbach, and Herbert Van de Sompel. Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library. D-Lib Magazine, November 2003, vol. 9, number 11.
<http://www.dlib.org/dlib/november03/bekaert/11bekaert.html>
- Belder93 Kurt de Belder. Electronic texts: A promise for humanities research. Academic Computing and Networking at NYU, vol 3/no.4, may 1993. available at:
<http://library.nyu.edu/research/french/text.html>
- Boonstra04 Onno Boonstra, Leen Breure and Peter Doorn. Past, Present and Future of Historical Information Science NIWI-KNAW report. Also Published in Historical Social Research/Historische Sozialforschung, Vol. 29 (2004), no2.
- Crane98 The perseus project and beyond. How building a digital library challenges the humanities and technology. D-lib Magazine, January 1998.
<http://www.dlib.org/dlib/january98/01crane.html> .
- Crane00 G. Crane (2000). 'Designing Documents to Enhance the Performance of Digital Libraries. Time, Space, People and a Digital Library on London', D-Lib Magazine 6 (7/8).
- Courtauld98 Courtauld Gallery Paintings: Illustrated Catalogue. A production of the Witt Library, Courtauld Institute of Art, London UK and Computer & Letteren, Univ. Utrecht. This is one of many such examples.
- Ditmarsch00 Hans van Ditmarsch: Knowledge Games. LLC Publications, Dissertation (DS) Series, UvA 2000. (<http://www.ilc.uva.nl/Publications/Dissertations/DS-2000-06.abstract.txt>)
- Doorn00 P. Doorn (2000). 'The Old and the Beautiful. A Soap Opera about Misunderstanding between Historians and Models', in: L. Borodkin and P. Doorn, *Data Modelling Modelling History. Proceedings of the XI International Conference of the Association for History and Computing, Moscow, August 1996.* Moscow University Press, 2-29.
- Eamon94 William Eamon. Science and the Secrets of nature. Books of secrets in medieval and early modern culture. Princeton UP, 1994.
- Fellbaum98 Christiane Fellbaum (ed.) Wordnet. An electronic lexical database. MIT Press, 1998.
- Fellbaum02 Christiane Fellbaum. On the Semantics of Troponymy. In Rebecca Green, Carol A. Bean and Sung Hyon Myaeng. The semantics of relationships. An interdisciplinary Perspective. Kluwer Academic Publishers, 2002.
- Gilchrist02 Alan Gilchrist. Thesauri, taxonomies and ontologies -an etymological note. jnl. of Documentation. vol.59.No1. 2003. pp7-18.
- Golfarb00 Charles F. Golfarb and Paul Prescod. The XML handbook. 2nd edition. Prentice Hall PTR, 2000.
- Hey03 Tony Hey and Anne Trefethen. The data deluge: an e-science perspective. In: F. Berman, A.J.G. Hey and G.C. Fox (eds.) Grid Computing: Making the Global Infrastructure a Reality. John Wiley & Sons Ltd, 2003. Also at :
www.rcuk.ac.uk/escience/documents/datadeluge.pdf
- Kircz00 J.G. Kircz and F.A.P. Harmsze. Modular scenarios in the electronic age. Conferentie Informatiewetenschap 2000. Doelen, Rotterdam 5 april 2000. In: P. van der Vet en P. de Bra (eds.) CS-Report 00-20. Proceedings Conferentie Informatiewetenschap 2000. De Doelen Utrecht (sic), 5 april 2000. pp. 31-43. Available via: www.kra.nl

- Kircz01 Joost G. Kircz. New practices for electronic publishing 1: Will the scientific paper keep its form. *Learned Publishing*. Volume 14. Number 4, October 2001. pp. 265-272. See: www.learned-publishing.org
- Kircz02 Joost G. Kircz. New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing*. Volume 15. Number 1, January 2002. pp. 27-32. See: www.learned-publishing.org
- Kirschner03 Paul A.Kirschner, Simon J. Buckingham Shum and Chad S. Carr (eds). *Visualizing Argumentation. Software tools for collaborative and educational sense-making*. Springer, 2003.
- Klijn03 Edwin Klijn (ed.) SEPIADES. Recommendations for cataloguing photographic collections. European Commission on Preservation and Access, 2003. See: <http://www.knaw.nl/ecpa/sepia/index.html>
- McCarty03 Willard McCarty: "Humanities computing". In: Miriam Drake (ed.) *Encyclopedia of Library and Information Science*, 2nd edn.,: Dekker, 2003, pp. 1224-35. Also via: <http://www.kcl.ac.uk/humanities/cch/wlm/index.html>
- McCarty03b Willard McCarty. Depth, markup and Modelling. CHWP A25. (Computing in the Humanities Working Papers) published. September 2003. http://www.chass.utoronto.ca/epc/chwp/chc2003/mccarty_b2.htm
- McCrank02 Lawrence J. McCrank. *Historical Information Science. An emerging Unidiscipline*. Information Today Inc. 2002.
- Moor99 Aldo de Moor. *Empowering Communities. A method for the legitimate user-driven specification of network information systems*. PhD thesis Tilburg 1999. Center Dissertation Series. Tilburg university.
- Nentwich03 Michael Nentwich. *Cyberscience. Research in the Age of the Internet*. Austrian Academy of Science Press 2003.
- Netwerk04 Netwerk TV1. AVRO. Onderwerp: Nieuwe Bijbelvertaling Presentatie: Karel van de Graaf . Verslaggeving: Koen van Groesen en Josefin Hoenders Uitzenddatum: Do. 8 april 2004, 20.30u. <http://www.netwerk.tv/index.jsp?p=items&r=netwerk&a=107028>
- Herwijnen90 Eric van Herwijnen. *Practical SGML*. Kluwer Academic Publishers 1990.
- Oliveira04 Edgard Costa Oliveira. Towards a new authoring environment: overview of some ontology-based systems. Submitted to the ELPUB2004 conference. www.elpub.net
- Snow59 C.P. Snow. *The Two Cultures*. The Rede Lecture 1959. With an Introduction of Stefan Collini. Cambridge Univ. Press. Canto Edition 1996.
- Steele96 Kenneth B. Steele. 'The whole wealth of thy wit in an instant': TACT and the explicit structures of Shakespeare's Plays. CHWP B2. Publ. May 1996. <http://www.chass.utoronto.ca/epc/chwp/steele>
- Talstra92 Schermen met Schrift. De combinatie van bijbelwetenschappen en computer geïllustreerd aan de tekst van Genesis 48. Inaugurale rede Prof. Dr E. Talstra Vrije Universiteit 4 juni 1992.
- Thaller98 Manfred Thaller. Warum brauchen die Geschichtswissenschaften fachspezifische Datentechnische Lösungen? Das Beispiel kontextsensitiver Datenbanken. *Computer in den Geisteswissenschaften. Konzepte und Berichte*, 7, 237-264, 1989.
- Vickery97 B.C. Vickery. Ontologies. *Jnl. of Inf. Science*, 23 (4) 1997, pp.277-286.

R2 {Institutional reports}

- Alf03 Alfa@net. A knowledge domain for Dutch history, language and culture. A proposal for NWO-Groot Investment 2003-2004
- Bijker03 W.E.Bijker (chair). Building the KNAW International Research Institute on e-Science Studies in the Humanities and Social Sciences (IRISS). KNAW 2003
- Catch04 Catch. A research programme for continuous access to cultural heritage. NWO, Draft March 2004.
- Cole03 Keith Cole, Kevin Schürer, Hilary Beedham and Terry Hewitt “Grid enabling quantitative social science datasets- A scoping study. Economic & Social Research Council / Economic and Social Data Service, 2003
- Den05 Beleidsplan 2005-2008. Vereniging Digitaal Erfgoed Nederland www.den.nl
- Dijk03 N.M.H. van Dijk (voorz.) Behouden Toekomst. Een advies met betrekking tot de toekomst van de diensten van het Nederlands Instituut voor Wetenschappelijke Informatie. KNAW 2003
- Dijken03 Koos van Dijken en Natasha Stroker. Naar een publieksgericht archiefbestel. 100bv Economisch onderzoek voor de publieke sector. Zoetermeer 2003.
- Fielding 03 Nigel G. Fielding: Qualitative research and E-social Science: appraising the potential. 2003. via: <http://www.esrc.ac.uk/esrccontent/researchfunding/esciencecentre.asp>
- Hoven00 Calculamus@human. Een voorverkenning naar de plaats van de geesteswetenschappen in de informatiesamenleving. Gerard Drosterij, Jeroen van den Hoven, Gert-jan Lokhorst, Jos de Mul, en Irma van de Ploeg. Centrum voor de Filosofie van de ICT. Geschreven in opdracht van de Adviesraad voor het Wetenschaps- en technologiebeleid.
- Keo04 Kenniscentrum Elektronische Overheid, a portal for all governmental initiatives and relevant documents: <http://www.elo.nl/elo/index.jsp>.
- NeS02 The National e-Science Centre Report 2001-2002
- Surf03 SURF Strategic Plan 2003-6 The Heart of the Matter: _____
<http://www.surf.nl/download/MJP-final.pdf>
- Plugge03 L. Plugge (ed.) De Vruchten plukken. Trends en Visie. Informatie-en communicatietechnologie voor het hoger onderwijs. SURF 2003
- PWC02 PWC Consulting: ICT gebruik in musea. Een internationale vergelijking 2002. In opdracht van het ministerie van OC&W. Almere 2002.
- Ike02 The UK e-Science programme. Annual report. year 1: April 2001 -April 2002. On: www.nesc.ac.uk/news/ukesciencereport02.html
- Voorbij97 J.B. Voorbij (voorz.), P.K. Doorn, L.J. Engels, H. Franken, D. van der Plas, R.J.H. Scha, A.W. Willemsen. 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. KNAW 1997.
- Woolgar03 Steve Woolgar: Social shaping perspectives on e-science and e-social science: the case for research support. A consultative study for the ESRC. no date, but most presumably 2003. Via: <http://www.esrc.ac.uk/esrccontent/researchfunding/esciencecentre.asp>

R3 [Websites]

- 0- SURF: www.surf.nl
- 1- Iconclass. <http://www.iconclass.nl/>
- 2- World Wide Web Consortium (W3C): www.w3c.org
- 3- Dublin Core: <http://dublincore.org/>
- 4- Digitaal Erfgoed Nederland: www.den.nl/standaard/mesometa.html
- 5- Rijksmuseum www.rijksmuseum.nl
- 6- Allard Pierson Museum <http://cf.uba.uva.nl/apm/>
- 7- Jewish museum Vienna <http://www.jmw.at/en/index.html>
- 8- Internet Archeology <http://intarch.ac.uk/>
- 9- W. Maurice Young Centre for Applied Ethics <http://www.ethics.ubc.ca/index.htm>
- 10- Virtual Reality Modelling Language: <http://www.w3.org/MarkUp/VRML/>
- 11- Semantic Web: <http://www.w3.org/2001/sw/>
- 12- Biblia Sacra; <http://www.bibliasacra.nl/>
- 13- De Bijbel in de Nederlandse Cultuur: <http://www.bijbelencultuur.nl>
- 14- Cultuurwijzer: www.cultuurwijzer.nl
- 15- Cave technology: http://www.supercomputer.nl/sc/watergraafsmeer_cave.html
- 16- SPSS: www.spss.com
- 17- Initiative for the Evaluation of XML Retrieval.
<http://inex.is.informatik.uni-duisburg.de:2003/proposals.html>
- 18- Text Encoding Initiative: <http://www.tei-c.org/>
- 19- Open Archive Initiative (OAI): <http://www.openarchives.org/>
- 20- OWL Web Ontology Language Overview:
<http://www.w3.org/TR/2004/REC-owl-features-20040210/>
- 21- Perseus digital library: <http://www.perseus.tufts.edu/>
- 22- Digidiva: <http://digidiv.amsterdam.nl/>
- 23- KB-Digital Preservation Research:
http://www.kb.nl/kb/resources/frameset_kb.html?kb/hrd/dd/dd.html
- 24- Electronic Metadatastructure for Endangered Languages Data (EMELD) <http://www.emeld.org>
- 25- Rosetta project: <http://www.rosettaproject.org/live>
- 26- XML stylesheet techniques, XSLT: <http://www.w3.org/TR/xslt>
- 27- EGA WEB ARCHIVE: An endangered languages documentation initiative.
<http://www.spectrum.uni-bielefeld.de/langdoc/EGA/>
- 28- AGORA, a web based framework for ethics education targeted at engineering students.
<http://www.surf.nl/projecten/index2.php?oid=114>
- 29- The European Network in Human language Technologies (ELSNET): www.elsnet.org
- 30- SourceForge: <http://sourceforge.net/>
- 31- Meertens Institute: <http://www.meertens.knaw.nl/>
- 32- Elsevier BV DTD for scientific papers:
http://www.info.sciencedirect.com/librarian_help/dtds/index.shtml
- 33- Science Direct: <http://www.sciencedirect.com/>
- 34- Spoken Dutch Corpus: <http://lands.let.kun.nl/cgn/ehome.htm>
- 35- Computer Physics Communications Program Library; <http://www.cpc.cs.qub.ac.uk/>
- 36- Nederlandse Taal unie: <http://taalunieversum.org/taalunie/>
- 37- Platform voor het Nederlands in Taal- en Spraaktechnologie:
http://taalunieversum.org/taal/technologie/platform_voor_het_nederlands_in_taal_en_spraaktechnologie/
- 38- Surfspot: <http://www.surfspot.nl/>
- 39- Mets: <http://www.loc.gov/standards/mets/>

List of Interlocutors

Dr. J.T.G. Arends

Leerstuurgroep Theoretische Taalwetenschappen. Universiteit van Amsterdam. E: j.t.g.arends@uva.nl

Kurt de Belder MA MLIS

Chief Division of Electronic Services, University Library Universiteit van Amsterdam. E: k.f.k.debelder@uva.nl

Prof. dr Jürgen. van den Berg,

Institute of Information and Computing Sciences. Universiteit Utrecht. E: jurgen@cs.uu.nl

Dr Martin P. Bossenbroek.

Head Expert services & collections, National Library of the Netherlands (KB). Den Haag. E: martin.bossenbroek@kb.nl

Dr Peter K. Doorn

Hoofd Afdeling Geschiedenis, Nederlands Instituut voor Wetenschappelijke Informatiediensten Koninklijke Nederlandse Akademie van Wetenschappen (NIWI-KNAW) E: peter.doorn@niwi.knaw.nl

Drs. Alice Dijkstra,

NWO Geesteswetenschappen, Den Haag. E: dijkstra@nwo.nl

Prof. dr Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany. E: gibbon@spectrum.uni-bielefeld.de

Drs. Gert Goris

Head Documentary Information Erasmus University, Rotterdam. E: Goris@ubib.eur.nl

Prof. dr M. Jeroen van den Hoven

Department of Philosophy, Faculty Technology Policy and Management. Delft University of Technology. E: m.j.vandenhoven@TBM.TUdelft.nl

Dr Mark Kas

NWO Exacte Wetenschappen Den Haag. E: kas@nwo.nl

Jaap Kloosterman

Director IISG international Institute of Social History, Amsterdam. E: jkl@iisg.nl

Steven Krauwer

Coodinator ELSNET (European Network in language and Speech), Utrecht Institute of Linguistics OTS, Utrecht University E: steven.krauwer@let.uu.nl

W.L.R. Mazeland

Remarco: business development in the legal environment. Driebergen. E: mazeland@inter.nl.net

Prof. Dr Willard McCarty

Centre for Computing in the Humanities King's College London E:

Willard.McCarty@kcl.ac.uk

Dr Michael Nentwich

Austrian Academy of Sciences, Inst. of Technology Assessment, Vienna Austria. E:

mnent@oeaw.ac.at

Dr Marc van Oostendorp

Meertens Instituut, KNAW, Amsterdam. E. marc.van.oostendorp@meertens.knaw.nl

Prof. dr Maarten de Rijke

Language & Inference Technology Group, Informatics Institute, University of

Amsterdam. E: mdr@science.uva.nl

Dr J. Peter Sigmond

Directeur collecties Rijksmuseum, Amsterdam. E: p.sigmond@rijksmuseum.nl

Jos Taekema

Directeur Vereniging Digitaal Erfgoed Nederland, Den Haag. E: jos.taekema@den.nl

Prof. dr Eep Talstra

Hoogleraar Oude Testament, Faculteit der Godgeleerdheid Vrije Universiteit

Amsterdam. E: e.talstra@th.vu.nl

Thorsten Trippel

Dept. of Linguistics and Literary Studies Bielefeld University Germany. E:

ttrippel@spectrum.uni-bielefeld.de

Dr J. (Hans) B. Voorbij

Institute of Information and Computing Sciences. Universiteit Utrecht. E:

hansv@cs.uu.nl

Dr. Peter Wittenburg

Max Planck Institute for Psycholinguistics, Nijmegen. E: peter.wittenburg@mpi.nl

E-based Humanities and E-humanities on a SURF platform

A report commissioned by SURF-DARE

Author: Dr Joost Kircz

Kircz Research Amsterdam

www.kra.nl

1 June 2004