ALLEA E-HUMANITIES WORKING GROUP REPORT

# GOING DIGITAL: CREATING CHANGE IN THE HUMANITIES

ALLEA

ALL European Academies

ALLEA E-HUMANITIES WORKING GROUP REPORT

# Going Digital: Creating Change in the Humanities


Dr Sandra Collins - Royal Irish Academy

Dr Natalie Harrower - Royal Irish Academy

Dr Dag Trygve Truslew Haug - Norwegian Academy of Science and Letters

Dr Beat Immenhauser - Swiss Academies of Arts and Sciences

Professor Gerhard Lauer - Union of German Academies of Sciences and Humanities

Professor Tito Orlandi - The National Academy of the Lincei

Professor Laurent Romary - DARIAH

Dr Eveline Wandl-Vogt - Austrian Academy of Sciences


A Position Paper outlining the ALLEA E-Humanities Working Group's recommendations on the current and emerging landscape and the key innovations and requirements for continued growth and excellence in the Digital Humanities.

# IMPRINT

Editor: Natalie Harrower
Design and Layout: Caitlin Hahn, Felicitas Soeiro
Printed and bound in Germany by: Druckerei WIRmachenDRUCK GmbH

Cover photograph: © bymandesigns / vectorstock.com

Legal Notice: This publication may be used for scientific purposes under citation of the source. The views expressed in this publication are the sole responsibility of the authors.

ISBN: 978-3-00-049483-3

ALLEA
ALL European Academies

© All European Academies, Berlin 2015

**ALLEA**, the Federation of All European Academies, was founded in 1994 and currently brings together over 50 Academies in more than 40 countries from the Council of Europe region. Member Academies operate as learned societies, think tanks and research performing organisations. They are self-governing communities of leaders of scholarly enquiry across all fields of the natural sciences, the social sciences and the humanities.

Independent from political, commercial and ideological interests, ALLEA's policy work seeks to contribute to improving the framework conditions under which science and scholarship can excel. Jointly with its Member Academies, ALLEA is in a position to address the full range of structural and policy issues facing Europe in science, research and innovation. In doing so, it is guided by a common understanding of Europe bound together by historical, social and political factors as well as for scientific and economic reasons.

**Contact:**

ALLEA Secretariat | c/o Berlin Brandenburg Academy of Sciences and Humanities
Jaegerstr. 22/23 | 10117 Berlin, Germany
Phone: +49 (0)30-3259873-72 | Fax: +49 (0)30-3259873-73
Email: secretariat@allea.org | Twitter: @ALLEA_academies | Website: www.allea.org

# Table of contents


Executive Summary.....5
Introduction.....6
Digital Humanities: the Opportunity.....9
The Life Cycle of Scholarly Data.....14
Digitisation and Encoding.....15
Data Analysis and Digital Curation.....17
Standards – Metadata, Vocabularies, Ontologies.....19
Linked Data.....21
Digital Humanities Tools.....23
Corpora.....25
Sustainable, Trusted Repositories and Infrastructure.....26
Digital Humanities Networks and Organisations.....31
Dissemination and Access to Scholarly Information.....34
Publishing and Licensing Data.....36
Long-term Digital Preservation.....38
Policy.....39
Training for Digital Humanities.....42
Recognition and Career Progression for Digital Humanities.....44
Conclusions and Recommendations.....45
Appendix: Resources and Further Information.....46

# Executive Summary

This report presents the recommendations of the ALLEA E-Humanities Working Group to ensure Europe retains a leading position in the Digital Humanities. The European Academies have an important contribution to make to debates regarding long-term cultural preservation and scholarship in the Digital Humanities. Our forward-looking review of the area leads to the following recommendations:

1. **Take a long-term view.**
Sustaining long-term archives of unique and important cultural artefacts is critical for Europe's leadership in Digital Humanities. Adopting best practice for infrastructures is essential.

2. **Encourage openness.**
Open Access to data and infrastructures enables enhanced research, research integrity and cost-effectiveness. Open Data needs to be adequately funded.

3. **Support your people.**
Training and career progression are essential to prevent the loss of the critical skills needed to retain our competitiveness in Europe. Data management roles need suitable recognition.

# Introduction

This report presents the work of the ALLEA E-Humanities Working Group.

The E-Humanities Working Group addresses the Digital Humanities, humanities data, open access for data and digital preservation, and sustained e-infrastructures and digital tools for the humanities. The European Academies constitute a unique pan-European knowledge base that is trusted, non-partisan and long-term. The Academies therefore have an important contribution to make to debates regarding sustained digital infrastructures and project-funded artefacts, long-term durable digital preservation, and the societal responsibility for the preservation of our digital cultural heritage.

In this report we identify a number of key enablers for Digital Humanities:

- Responsible open access to humanities data
- Sustained and openly accessible research infrastructures for the humanities
- Long-term durable digital preservation to prevent digital obsolescence
- Support for the preservation of our digital cultural heritage
- Recognition for the scholars and practitioners that advance these areas.

Humanities data can be rich and complex, non-standardised in format, without common or consistent metadata and ontologies, and subject to complex rights issues. Consensus and best practice regarding digitisation and metadata standards for common usage that still retain the richness of different disciplines and data types could enable open access to Humanities data and enriched scholarship, leveraging the treasures in archives, repositories and libraries across Europe.

I would like to thank the Working Group members who have contributed tirelessly and from their deep expertise to this report:

- Dr Natalie Harrower - Royal Irish Academy
- Dr Dag Trygve Truslew Haug - Norwegian Academy of Science and Letters
- Dr Beat Immenhauser - Swiss Academies of Arts and Sciences
- Professor Gerhard Lauer - Union of German Academies of Sciences and Humanities
- Professor Tito Orlandi - The National Academy of the Lincei
- Professor Laurent Romary (Invited Expert Member) - DARIAH
- Dr Eveline Wandl-Vogt - Austrian Academy of Sciences

Thanks are also due to the staff and President of ALLEA for their support and guidance, without which this report would not be possible.

---Dr Sandra Collins (Chair, E-Humanities Working Group) - Royal Irish Academy

## ALLEA - European Federation of Academies of Sciences and Humanities

ALLEA, the All European Academies, was founded in 1994 and is a federation of almost 60 Academies of Sciences and Humanities from more than 40 countries in the Council of Europe region. Member Academies are self-governing communities of scientists and scholars across all fields of the natural sciences, the social sciences, and the humanities and operate as learned societies, think-tanks, grant givers and research performing organisations.

ALLEA's objectives are to promote the exchange of information and experiences between Academies, offer European science and society advice from its Member Academies, strive for excellence in science and scholarship, for high ethical standards in the conduct of research, and for independence from political, commercial and ideological interests.

ALLEA seeks to contribute to improving the framework conditions under which science and scholarship can excel. It is guided by a common understanding of Europe bound together by historical, social and political factors as well as for scientific and economic reasons.

ALLEA's "policy for science" work addresses the broader framework conditions for science and research in Europe and beyond. ALLEA directly involves its Member Academies in its deliberations, relying on the expertise of the Academies' leading scientists and scholars as well as its partner organisations. ALLEA advice reaches a wide range of decision-makers and stakeholders in the science policy arena as well as among the interested public.

ALLEA's positions are elaborated through permanent and issue-focused expert Working Groups and reflect on the societal, technological and environmental challenges that science faces, thereby proposing the steps necessary to maintain and expand a vigorous and rigorous science base in Europe.

ALLEA's main focus areas include the European Research Area and Horizon 2020, Digitisation and Research Infrastructures, Intellectual Property Rights and Open Access, Science and Ethics, Science Education, and the Social Sciences & Humanities.

## The ALLEA E-Humanities Working Group

ALLEA E-Humanities Working Group was established by the President of ALLEA, Prof. Günter Stock, at a meeting in the Berlin-Brandenburg Academy of Sciences and Humanities on 5 November 2012.

The E-Humanities Working Group is charged with identifying and raising awareness for priorities and concerns of the Digital Humanities, contributing to the Open Access agenda from a Humanities and Social Sciences perspective, and building consensus for common standards and best practices in E-Humanities scholarship and digitisation.

The European Academies constitute a unique pan-European knowledge base that is trusted, non-partisan and long-term. The Academies therefore have an important contribution to make to debates regarding sustained digital infrastructures and project-funded artefacts, the achievement of long-term durable digital preservation, and the societal responsibility for the preservation of our digital cultural heritage.

The E-Humanities Working Group was founded on 5th November 2012, and met throughout 2012-2014, both in dedicated group meetings, but also in the context of contributing to major European conferences including the ALLEA Scientific Symposium in 2012 and the ESFRI Facing the Future Conference in 2013.

The Working Group includes members from the following European Academies:

- Austrian Academy of Sciences
- National Academy of the Lincei
- Norwegian Academy of Science and Letters
- Royal Irish Academy
- Swiss Academies of Arts and Sciences
- Union of German Academies of Sciences and Humanities

# Digital Humanities: the Opportunity

Humanities are a self-evident part of modern societies. In the 19th century, nation building was heavily based on ideas of culture – each nation having its own uniqueness through language, literature and culture. The Humanities were established as an academic discipline to analyse, reflect, and practice culture and cultural heritage. The humanities and modern societies are therefore siblings. Although disciplinary approaches and even national traditions shape the way humanities are established, a common characteristic of the humanities are their historical and hermeneutical methods.

For more than half a century quantitative and statistical methods were in the background of scholarly methods. With the rise of computers, internet and big data, methods change quickly and new, more quantitative approaches are becoming increasingly important. Millions of digitised books and objects, fast access, huge storage, and new ways of digital collaboration and sharing transform culture and the humanities. 'Digital Humanities' is an umbrella term that circumscribes this change in scholarly objects and methods. In this broad sense, Digital Humanities is not the 'next big thing' in the humanities nor is it another turn. Instead, it simply indicates that the humanities as we know them have changed in the way they do research, how they teach, and how they are institutionalised. "Digital Humanities" covers a great number of activities in research, teaching, and cultural production, using computers and the internet as new technologies for scholars and the general public. In short, Digital Humanities fosters and broadens the methods and research opportunities in the humanities.

When computing in the humanities started, the first electronic editions like Roberto Busa's 'Corpus Thomisticum' or Jean-Claude Gardin's cataloguing of archaeological objects dominated the still small field of computer-based approaches in the humanities. In the 1990s, catalogues, dictionaries and encyclopaedias quickly became a growing part of larger retro-digitisation programmes. With an increasing number of tools at the beginning of the 20th century, computing in the humanities means more than simply finding strings of words or co-occurrences, collocates or concordances. Tools for linguistic research, or for research in musicology or art history, make computer-based methods widely available for most scholarly disciplines. Since approximately 2000 the new name 'Digital Humanities' has indicated these change, and endorses the achieved level for computing in the humanities as being a regular part of the humanities. To be more precise: many fields, such as archaeology, are now already exclusively digital archaeology, and early debates about the value of Digital Humanities have become a moot point relegated to the 'growing pains' of disciplinary history.

All humanities disciplines are being deeply influenced and even changed today by the opportunities offered by digital devices; in fact, they have prompted a revolution in the behaviour of scholars. This revolution is deeper than in the natural sciences, which were already equipped from their modern foundation to treat numerical and highly formalised data. This revolution may be welcome, but it needs to be approached with caution because of the tendency to ignore the methodological impact of digital instruments on the traditional approach to study in the humanities. It is correct to say that scholars are confronted not only with new instruments, but with a new language by which their data are represented and their contributions disseminated. This passage cannot be left to collaborators specialised in technology, but must become part of the competence of every humanist.

Other similar problems arise from the fact that the humanities are traditionally considered as uniform and coherent, but the fact that they cover a number of disciplines with different foci and methodologies means that the humanities are actually heterogeneous. For example, archaeology, linguistics, librarian and archival sciences, music, and cultural heritage studies are quite distinct from one another and have developed specific methodologies that require different special digital tools. On the other hand, some typical digital technologies are used in a way that crosses these boundaries, and their adjustment to humanistic demand may be much improved by a dialogue between different disciplines. Word processing and markup standards are as equally useful in literature as in history or epigraphy, and database management systems are employed in art history as much as they are in philology, but their use is not exactly the same. Digital Humanities drive the scholarly tradition towards a more uniform approach or an approach where 'two cultures' – humanities on the one side, and computing on the other, is no longer constructive.

Digital Humanities opens access to resources, revealing data that may have been enclosed in a special collection. With digital versions and tools, corpora can be right at the researcher's fingertips. The digital edition of Mozart's complete works, for example, offers the proof referenced by many theses about his compositions to researchers worldwide. Each day, more than 100,000 music lovers use the digital Mozart edition with its full texts, critical comments, libretti, letters, and documents. GIS data of archaeological excavations could be compared relatively easily with other data compiled at different digs. X-ray and UV pictures of famous paintings reveal previous versions of the same painting beneath the visible surface. Computed Axila Tomography scans even reveal evidence of the wood grain used in the original panels. Endeavours like this are offered by digital editions like 'Universal Leonardo', the complete works of Leonardo da Vinci. All in all, Digital Humanities research influences the reliability of scholarly findings as well as the validity of hypotheses and opens up new avenues of inquiry.

DH initiatives offer scholars access to cultural heritage in unconceivable breadth and depth. Millions of books are only a few clicks away. The web portal 'Gallica', run by the Bibliothèque Nationale de France, offers more than 1.5 million books, documents, periodicals, images, sound recordings and scores in the French and European tradition since 1500, and 100,000 documents per year are added to the existing collection. 'Eighteenth century collections online', to mention just another example, unlocks about 200,000 English-language and foreign-language titles printed in the United Kingdom in

Left and above: Leonardo da Vinci, Madonna of the Yarnwinder (The Lansdowne Madonna), 1501-07, Selecting a type of scientific analysis/Ultraviolet analysis. Screenshots of Universal Leonardo website, available at www.universalleonardo.org

© University of the Arts, London 2015

the 18th century. And the 'Verzeichnis Deutscher Drucke des 18. Jahrhunderts' will digitise about half of all the books printed in German-speaking countries from 1700 to 1800. Research is no longer solely bound to exploiting the canon of 'great books' and famous, widely-known art. Where standard literary histories mention only one German novel for the year 1809 —Goethe's 'Wahlverwandtschaften'— a quick search at Google Books shows nearly a hundred mostly unknown novels published in the same year.

Research methods are not only restricted to historical and hermeneutical approaches. Quantitative and more or less statistical approaches widen the inventory of scholarly methods. Different statistical measures, for example, reveal how differently we speak about social problems like unemployment in contrast to natural events like earthquakes. In sum, the so-called gap between humanities research and more typically quantitative or 'scientific' approaches is narrowing, and the transformation of scholarly tradition through the dissolution of disciplinary boundaries is well underway.

Still, many of these new methodological opportunities are not well developed. Access to cultural heritage materials is restricted by copyright laws that recall the traditions of 19th century thinking. Projects run by Academies long before the digital age began are bound to contracts which do not allow data sharing. Shared data is still not standard in scholarly nor in scientific research practice. New methods are seldom taught in the regular course of humanities courses, and when they do surface they are generally confined to special summer schools or provided through e-infrastructure initiatives such as DARIAH or CLARIN. The debate also remains around whether or not quantitative and computational methods are complementary to established historic-hermeneutical methods. It is therefore necessary to encourage scholars to step into the more transparent research paradigm that Digital Humanities now offers, and to be critical about what is required to develop the computational aspect of the humanities. The Research Data Alliance, funded by the European Commission, the Australian Government and the National Science Foundation in the United States of America, is one body making headway in data sharing, and its outputs should be encouraged for adoption.

Digital resources offer a whole new set of possibilities for the humanities since they can be explored by computational means. Although computers are not able to read and understand as humans do, they are able to deal with large amounts of data in a fast and standardised way. For example, computers are able to analyse data like text, visual arts or music. A case in point is authorship attribution. Traditionally, this relies on the scholar's judgment of the style of a certain text (in addition to real-world correlates such as references to historical events). Even at the pre-digital stage this was sometimes studied using methods that are essentially computational, e.g. by counting occurrences of words and analysing the patterns statistically by

hand. Computers offer obvious advantages, not just by counting quickly, but by eliminating the need to guess in advance which patterns will be significant, since the computer can extract significant patterns more or less automatically. As the seminal work by John Burrows on text analysis has shown, the simple frequency of word use identifies authors and genres. Similar methods can be applied across a whole range of problems in the humanities, since pattern recognition is a fundamental method in many approaches beyond authorship attribution, such as textual criticism and language relationships. Nor are the methods limited to textual data - more advance methods like visual analytics unlock comparative studies on a large scale.

The output of such methods relies crucially on the quality of the data input, and here projects run by Academies are relevant. The resources that are created in Academy projects lend themselves particularly well to this kind of computer-based study as they are painstakingly precise in their coding of data. Statistical methods sometimes rely on accessing large quantities of data, which has the side-effect of generating noise in the input data; however, for many problems in the humanities, the amount of available data is very limited and can't be easily expanded. When we study the authorship of a particular text, or the relationship between different medieval manuscripts, or the authenticity of a composition said to be by Mozart, we cannot enlarge the amount of data available, so it becomes even more important that the data is correct, authoritative and well-structured.

Given the leading role that Academies have played in providing such cultural data in digital formats, they should also take the opportunity to lead the way in the use of these resources, not only as advanced books, but by focusing on the genuine new possibilities that they open up.

The Meeting Room of the Royal Irish Academy, Dawson St, Dublin

# The Life Cycle of Scholarly Data

The advent of digital methods in the humanities means that scholarly data has a life of its own. Formerly, scholarly data would in the best case be available in archives or in printed books, typically alongside the research that it went into. Nowadays, scholars can and should make their data available as independent resources that can be copied, modified, enriched and redistributed. It is this sharing culture that gives most of the added value to digital methods, for it leads to transparency, reproducibility and incremental progress in fields where this used to be difficult. But there are also challenges: sharing requires strict standards for data formats, good practices for attribution and the availability of digital archives for long-time preservation.

Since the life-cycle of scholarly data is therefore crucial to the Digital Humanities, we have structured our report around it. We start with the digitisation process, which is fundamental to data that is not born digital – be it heritage texts, the results of archaeological excavations, or recorded spoken material. Next, digitised and born-digital data must be standardised and linked to other resources, and we discuss best practices in this field. We then move on to corpora – structured data collections – and the tools that can be used to analyse them. Finally we come to the infrastructures that are needed to sustain, curate and preserve these data collections and disseminate them. None of this can exist without dedicated scholars and archivists, who must be trained and given proper recognition for their work.

Our report is linear but the life of scholarly data is cyclic. A piece of a New Testament manuscript may be digitised in one project. If the digitised text is standardised and/or well documented, it can then be reused in another project and for example assigned a morphological analysis. Another project may be more interested in how the discourse is structured, and yet another project may work on how that manuscript's rendering of a particular passage differs from those of other manuscripts, or even translations into other languages. And a later project may attempt to integrate all this research into a single resource. For this complex chain to work, each project must pay careful attention to each step in the data life-cycle. Doing this properly requires more dedicated resources than merely uploading some documents to a webpage, but the results of these dedicated efforts are worth it, because they enable the Digital Humanities to explore areas that were previously unreachable.

# Digitisation and Encoding

Most texts existing in digital form are born digital. They may be investigated and evaluated in many respects (and this is widely done already), but we can presume that, in comparison to texts first produced on paper, they are close to the way their author intended them to be. Texts that were produced in another form and then transferred to digital form pose other challenges, as it is not always clear how closely the digital form corresponds to its original.

Digitisation of texts is achieved in two different ways:

- visual representation through digital camera or scanning procedures;
- encoding of the written text as a sequence of alphanumerical entities, eventually accompanied by an encoded textual description of the material aspects of the texts, such as layout, by means of tags.

These two ways are quite different in terms of the technology used and in the properties of the results. In any case, both raise the same question: how much of the information contained in the original has been accurately represented in the digitised version? What we call 'text' is in fact a multi-level and complicated communication phenomenon, which at its birth and conclusion, e.g. in the mind of the author and of the reader, is something immaterial and ideal in the Platonic sense. In the process of communication it becomes the material representation of the mental message, through voice or script (or otherwise), produced by the author in order to convey the message to other people, or in any case to record it for the future.

Encoding is generally understood as the passage into digital form of a text considered as a sequence of alphanumeric and auxiliary characters. Encoding should include both the denotational aspects of the linguistic expression (e.g. the words on the page) as well as the connotational ones – their position on the written surface, comparative size, underlines, abbreviations, reference marks, colour, and so on. These aspects of the original text are required to provide a complete picture of the text's semantic signalling and should not be left out of the digital version file.

This is why tags have been conceived, as a means to insert as unobtrusively as possible paratextual information in a digitised text. The first standard proposed was sgml, followed by more explicit xml and then by its implementations, e.g. html, and TEI. This way to obtain in electronic format a text enriched by relevant paratextual information should be recommended by academic institutions. But the implementations named above are not exempt from problems, which should be accurately discussed and studied in the future.

## The Canterbury Tales edition by Peter Robinson

Canterbury Tales Project:
http://www.petermwrobinson.me.uk/canterburytalesproject.com/index.html

The (ongoing) Canterbury Tales project, led by Peter Robinson, is an example of an excellent achievement in the realm of digital editions.

The Canterbury Tales digital edition is based on a full-text transcription of original texts into electronic form, and this transcription is based on explicit, declared principles. The CD-ROMs include all the transcripts of the fifty-eight witnesses, images of all the pages of text in these manuscripts, the spelling databases developed as a by-product of the collation, collation in both 'regularised spelling' and 'original spelling' forms, and various descriptive and discursive materials. As such, it presents a mass of materials which an editor might use in the course of preparing an edition.

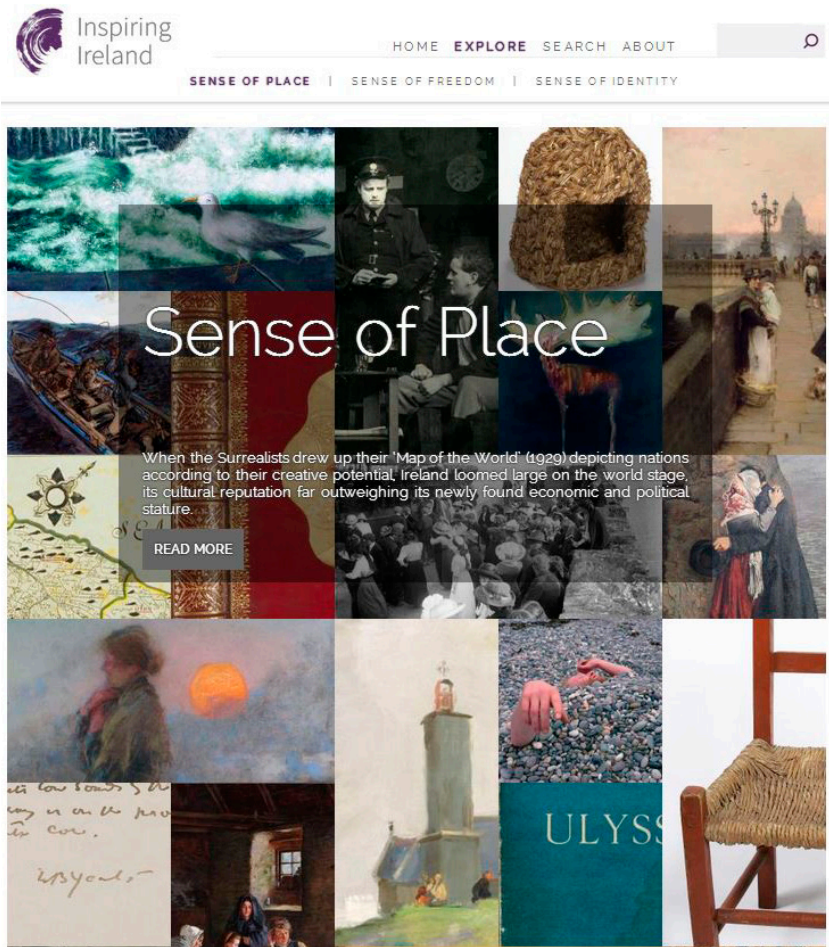
The edition is also an example of how the use of computer-assisted analytical methods may restore historical criticism of large textual traditions as a central aim for scholarly editors. The transcripts and images of the many versions of any one tale with collations and analyses offer readers the opportunity to efficiently check the stability of the text at critical points. By inviting exploration rather than hiding revision history through editorial decision-making, such editions might help us all to be better readers. The new technology has the power to alter both how editors edit, and how readers read.

As an editorial project generating substantial quantities of transcribed text in electronic form, the Canterbury Tales Project adopts an open transcription policy, modelled on the copyright licensing arrangements developed by the Open Software Foundation. It is important to note that this policy does not mean that institutions and individuals give up all copyright control. The originators of the transcripts still retain this, and so can still (where possible) make commercial arrangements for their publication and prevent inappropriate use. What it does mean is that the copyright holders assert that the transcripts may be freely downloaded, used, altered and republished subject to certain conditions. In short, the republication must be under the same conditions; all files must retain a notice with them to this effect; permission must still be sought for any paid-for publication).

In the case of born-digital data, the challenges are not in how to make it digital, but in how to make data persistent and interoperable. The internet itself is a born-digital collection of cultural data. But there is nothing more fluid than the internet and any internet researcher faces the problem of how to grasp the flow of data. The 'Internet Archive' initiative by Brewster Kahler is the first attempt to get a clearer picture how the internet evolves over time, and other projects have emerged to collect, organise and preserve various aspects of data on the internet, such as the ARCOMEM EU project, which looks at the social web and memory institutions, or the Library of Congress's initiative to archive all of Twitter. Such initiatives, which collect, curate, and preserve born-digital content, will become more and more central for scholarly research.

# Data Analysis and Digital Curation

As argued above, creating digital resources for the humanities is not merely a prerequisite for the humanities of the future, it is an important scholarly activity in itself. The simplest forms of digitisation, images of existing documents in JPG or PDF format, are themselves mostly useful as a way of making objects accessible to researchers. However, such data cannot in general be further processed by a computer without additional information; at a very basic level, digitised data require descriptive metadata.

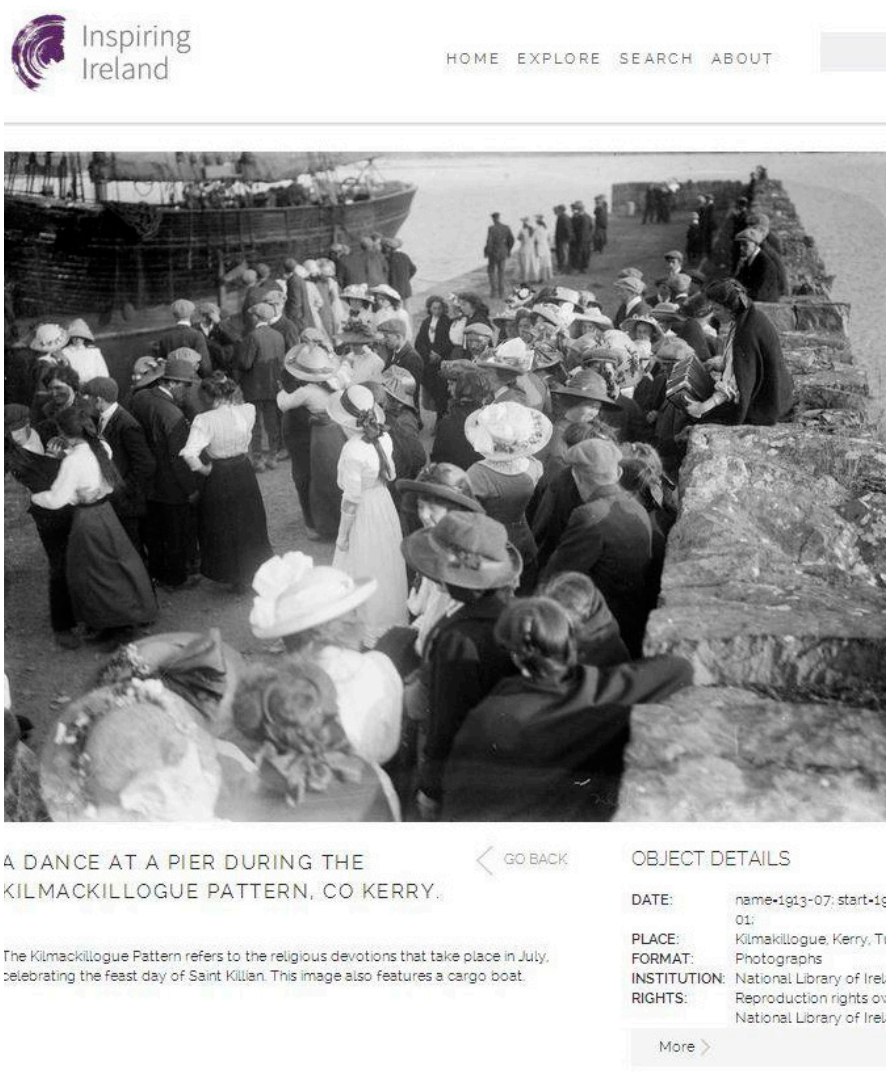



To be more useful to researchers, digitised data needs to be structured. This involves encoding an understanding of the data: at the most basic level, this may be a transcription indicating that a piece of black ink in the picture file is an instance of a particular letter in a text, or perhaps a particular note in a piece of music. But more advanced structures can increase the usefulness of the data immensely. We may want to know what word the letter is part of - not necessarily a trivial decision if the script in question does not distinguish word boundaries as our modern European alphabets do. We may also want to know what sentence that word was part of, and if the sentence belongs to a play, for example, whether the sentence occurs in scene descriptions or in the spoken dialogue, and if so, which character spoke it. At a deeper level we may want to know where the sentence belongs in the structure of the text, either linguistically or content-wise, perhaps according to well-established structural typologies. Providing such analyses is a scholarly activity akin to time-honoured work in the humanities such as the creation of textual editions, and deserves to be recognised as such.

Encoding such analyses in a computer-readable way allows humanities researchers to collaborate and vastly expands what a single

researcher can achieve on her or his own. But to build these big infrastructures, it is often necessary to combine earlier work from different sources and to extend it, which requires interoperability at both the technical and the legal level. Data must conform to extant standards and formats so as to be readily usable by others, and copyright/licensing must allow other researchers to modify and redistribute it.

When it comes to visual data in digital formats (such as an image of a painting) search and discoverability is only made possible through the metadata that accompanies the digital object. Digitised cultural objects require curation, which includes rich, structured data for context and scholarly access. Metadata includes both authoritative descriptive content about the object, which is often written by art historians, as well as technical information about the provenance, format and quality of the digital object itself (whether this is a born-digital object, such as a digital photograph that contains EXIF data, or a digital surrogate of an existing material object, such as a photograph of a painting or sculpture). Digital curation adds to an object's discovery for scholars, and thus provides new angles for research. An example of a recent project that both extends access to cultural objects and also develops the scholarship around those artworks in the process is the Inspiring Ireland website, which contains a cross-searchable selection of digitised cultural artefacts, organised by theme and enriched by gallery curators and historians, from eight of Ireland's national cultural institutions. This project is notable because it uses a Digital Humanities approach to potentially disparate objects, increases public access, and also preserves the images in a trusted repository.




Previous page and above: Inspiring Ireland screenshots, available at www.inspiring-ireland.ie.

# Standards – Metadata, Vocabularies, Ontologies

The digital age is the age of standardisation. Whenever we talk of interoperability, or of the integration of heterogeneous data, the use of standards is required for any further success. Data can only be transformed into information and information into knowledge if standards are shared. Libraries are one of the major institutions to make the most out of data. They manage metadata standards that describe resources carefully and connected pools of data. The Worldcat connects 7,000 world libraries with millions of holdings in a common metadata standard, which enables users to explore long-term cultural trends. An example is the entry on Johann Sebastian Bach, which shows scholarship by and about Bach, in hundreds of thousands of publications around the world, and provides a visualisation timeline to quickly note trends.

Vocabularies and ontologies are a means to define concepts in a structured manner, and characterise the relationships between these concepts, i.e. to define a data model. An example of such a data model is Europeana's data model, formalised using the Resource Description Framework (RDF). Vocabularies and ontologies can be used to relate data to standard domains of knowledge, facilitating integration of different datasets, organising knowledge and making sense out of a diversity of data.

Many standards have been developed. TEI, the Text Encoding Initiative, is the one of the best examples of how standardisation could work on an international level. And with good reasons, initiatives like MEI, the Music Encoding Initiative, follow this example.

Standards are documents informing about practices, protocols, artefact characteristics or data formats that can be used as reference for two parties working in the same field of activity to be able to produce comparable (or interoperable) results. Standards are usually published by standardisation organisations (such as ISO, W3C or the TEI consortium), which ensure that the following three requirements for standards are actually fulfilled:

- Expression of a consensus: the standard should reflect the expertise of a wide (possibly international) group of experts in the field
- Publication: the standard should be accessible to anyone who wants to know its content
- Maintenance: the standard is updated, replaced or deprecated depending on the evolution of the corresponding technical field

Standards are not regulations. There is no obligation to follow them except when one wants to produce results that can be compared with those of a wider community. Ideally, a good standard reflects the work of a

relevant community and is maintained by the appropriate body. Because there is no obligation to use a given standard, it is essential to provide potential users with a) awareness about the appropriate standards and the interest to adopt them, and b) the cognitive tools to help them identify the optimal use of standards through the selection and possibly customisation of a reference portfolio.

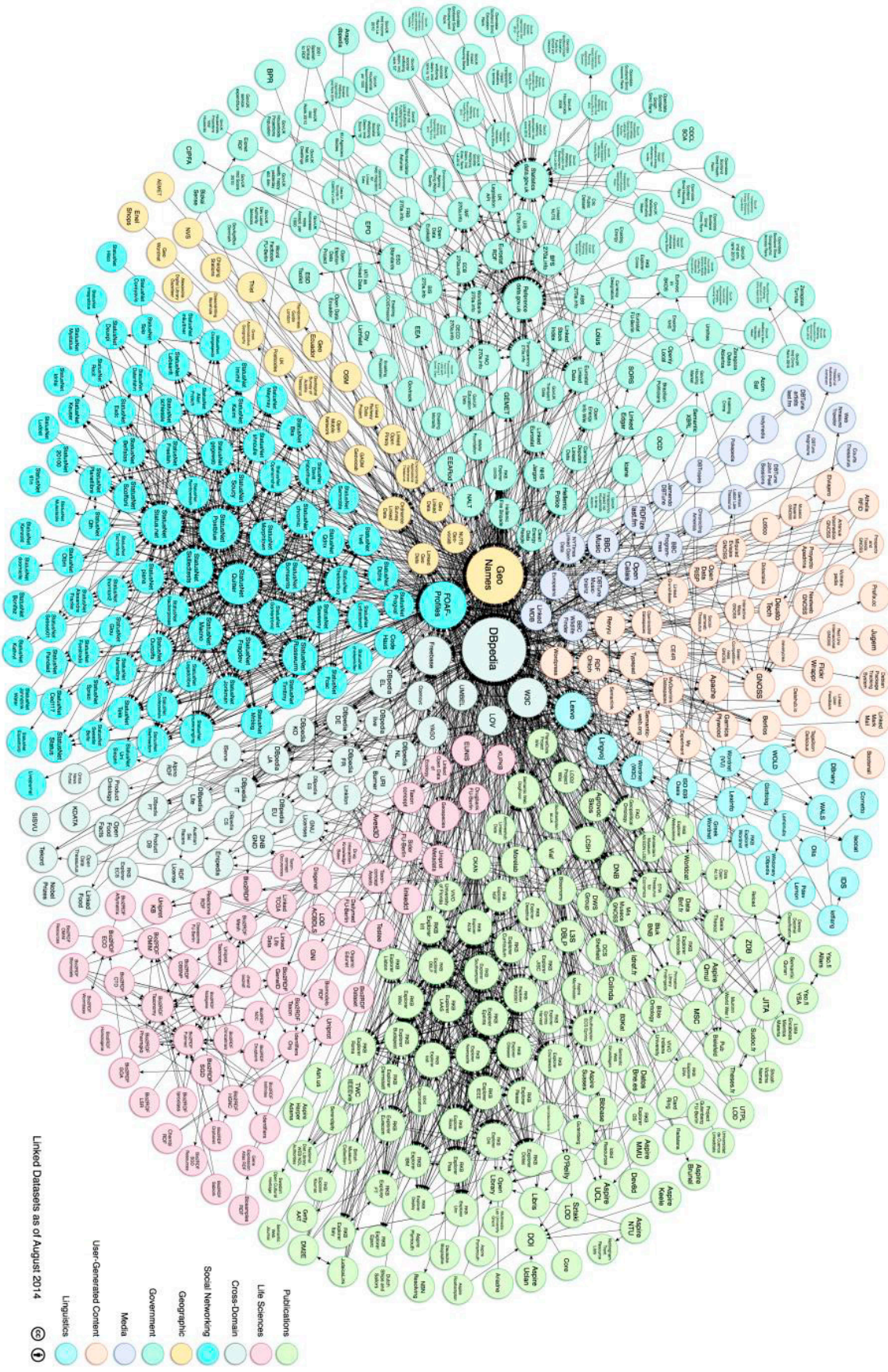
Above: Bach, Johann Sebastian, Publication timeline of 133,340 works in 281,768 publications in 18 languages and 1,403,223 library holdings from 1685 – 2013. Screenshot of Worldcat Identities, available at http://www.worldcat.org

# Linked Data

Linked Data refers to data published on the Web following a set of principles designed to promote linking between entities on the Web. An essential requirement to enable this linking is that each entity (for example a place name or personal name) is given a unique identifier, generally in the form of a Uniform Resource Identifier (URI). Having determined these URI identifiers, Linked Data reuses other data models such as the Resource Description Framework (RDF) to specify the links, and their type, between two URIs. Linked Data technologies such as identifiers, data models, knowledge representation languages and ontologies allow us to integrate humanities and social sciences data and increase its visibility on the Web.

Linked Data technologies are being used by universities and cultural institutions across the world to improve the searchability of data, content discovery, data integration and re-use, and to improve education and learning technologies. The Linking Open Data diagram on the following page shows the extent of existing data sets and their connections as of August 2014, with humanities data starting to play a more and more significant role. For example, Europeana is piloting Linked Data to structure and represent cultural heritage data for additional functionality, contextualisation, exchange and re-use by external communities according to their own needs. Projects such as the EU FP7-funded LATC: LOD Around The Clock and the European Commission-funded Digitised Manuscripts for Europeana (DM2E) support and improve the development of tools and better infrastructure for Digital Humanities, sciences and multi-domain Linked Data. As research becomes more data-intensive, collaborative, and multi-disciplinary, the ability to link our data and datasets is crucial for the integration of research between and across sectors, providing additional context, deepening understanding and opening up new research questions.

Linked Data is set to become a formal standard of the W3C soon and will be a fundamental part of the future of the Web. It will be followed by a growth in the development of tools and applications to create and use Linked Data for research and across the Web. The use of well-documented, best practice standards such as Linked Data ensures the future-proofing of research data, the central position of research in the future Web environment and the ideal conditions under which science, humanities and scholarship can excel.



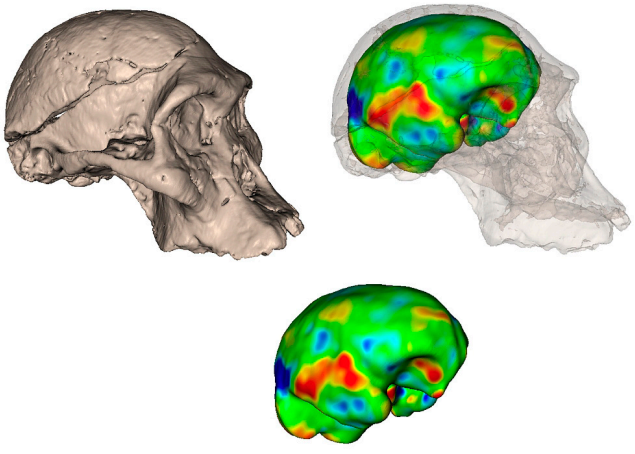

Above: inking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/ License: CC-BY-SA

# Digital Humanities Tools

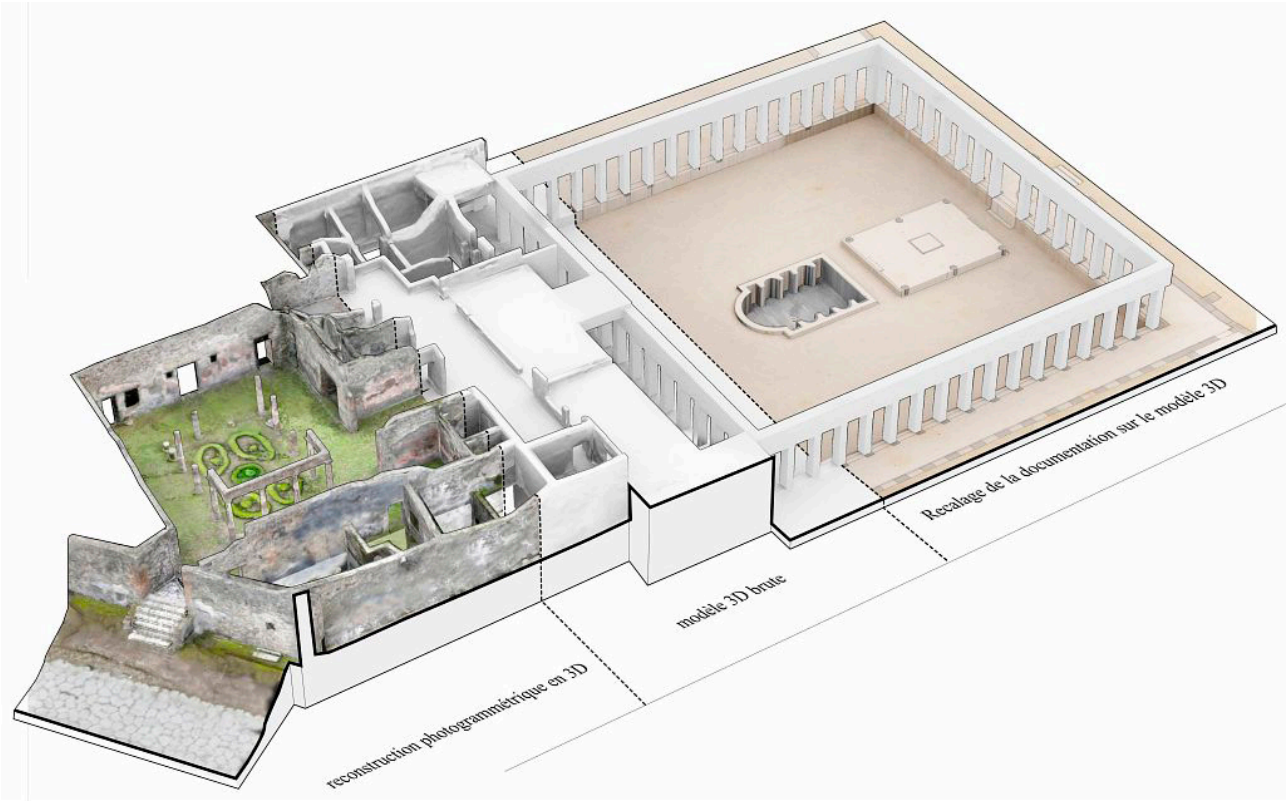
In recent years a growing number of tools open more and more research opportunities for humanities disciplines. Stylometry and phylometry, topic modeling and sentiment analysis, machine learning procedures, 3D modelling, rti photography, geo-referencing or social network analysis – to mention a few of them – demonstrate the increasing research opportunities. Stylometric approaches for example, based on the statistical analysis of 'little words', are able to distinguish between literary epochs and offer new ways of practicing literary historiography.

Comparable to former development in areas like computational physics, chemistry, or biology, Digital Humanities now could broaden the research opportunities for disciplines in the humanities by integrating quantitative and formal methods into the portfolio of methods. Open source tools like Gephi or Rapid-I, Voyant or Atlas.ti and non open tools like n-gram viewer – to name only a few – reveal the potential for scholarly research beyond hermeneutical approaches. A number of papers in journals like 'Science', which was traditionally not the domain of the humanities, show what is already possible. To give but one example: Digital Humanities facilitates 3D reconstructions, combining former excavations and old measurements with new photogrammetrical methods to enable a better understanding of ancient culture.

Therefore it is quite clear how Digital Humanities will become an integral part of humanities in general by the use of already developed tools. The humanities does not need support for new methods – these methods are already in practice, and lists of tools provided by DARIAH or info.clio help scholars find the right tools for the specific research question they have in mind. But these tools and methods need bridging support, i.e. teaching through workshops, summer schools, and integration into curricula to ensure strong methodological progress. Importantly, scholars and institutions also need a sustainable infrastructure for storing raw and processed research data, for operating and preserving tools, and for disseminating results.



Above: INRIA-CDR00046-0193



Above: Pence, J. et al.: 3D reconstruction of the Villa of Diomedes, Pompei, based on photogrammetry and old maps (INRIA-0133-056 and INRIA-0133-046)

# Corpora

The humanities think more or less in categories of canon and editions, mostly editions of great works. But the digital age is different in respect to editions. They are part of larger corpora. And the switch from thinking in categories of editions to categories of corpora is essential. In the digital age the limits of the printing age are mostly gone. Whether it is a picture or a sound, a fragment of text or a sculpture, in the digital world all are bits and bytes. The technical difference between 3D-Scans and TEI annotated texts, music files and architectonical plans, is very small, but in the way we document these files, their descriptive 'metadata' is not standardised, and it's this information that contextualises the digital object. Standards of interoperability are therefore of increasing importance for translating between all the different digital resources. The Europeana metadata aggregator makes visible how different digital formats could be integrated into one (meta-) library. A variety of resources could integrate cross-domain and cross-media content in one corpus. The 'European Holocaust Research Infrastructure (ERIH)' is a good example. To identify all names of the six million murdered Jews, the project makes use of heterogeneous sources, including tax records, deportation lists, diaries and personal audio testimonies.

A few years ago no one could imagine the sheer number of today's already digitised objects and texts and the fast growing number of born digital material. Google Books, with its estimated 30 million retro-digitised books (out of a total of approximately 130 million printed books in existence) is only the publicly known side of large scale cultural corpora. The Europeana library gives access to more than 10 million digital objects, for example items like Leonardo's 'Mona Lisa' or Mozart's music. And billions of fan fiction works are constantly published via the internet, not to mention the 100 hours of video uploaded every minute to YouTube or the 540,000 tweets sent per minute. To build corpora out of these collections and libraries offers a wider view of cultural processes, and perhaps of what one might call cultural evolution.

However, the ability to create such corpora depends heavily on the conceptualisation of cultural heritage as a corpus and not as a canonical edition. This intellectual shift will change the way cultural resources are made available in the near future. They could and they should be part of a global network of resources, shared by humankind. And it is more than just words that the digital age could make available. However, development appears to be happening in the opposite direction, with an oligopoly of companies owning more and more of our collective cultural heritage. It is a major task to 'free' our public heritage from private ownership. Thinking in categories of corpora is one way to approach a reframing of culture.

# Sustainable, Trusted Repositories and Infrastructures

As more and more data are born digital or digitised it is vital to implement and sustain trusted digital repositories that will maintain the records of the past and present – whether that be text, audio, image, moving image or multimedia – for wide access and re-use, for now and for future generations. With regard to Research Infrastructures (RIs) for the Digital Humanities, we primarily mean Digital Repositories, Archives and Databases of data relevant to the Humanities, and also digital platforms for discovery, mining, analysis, visualisation, curation and exhibition, as well as digital services, apps or tools. A first attempt has been made by the design phase of a Humanities Data Centre, started in 2014 (http://www.humanities-data-centre.org/). The data centre will ensure the long-term preservation of and access to research data across the range of humanities disciplines. Furthermore, it will provide tools and APIs for virtual research environments as well as counsel and training.

While the term 'research infrastructures' is used as a generic term for several different facilities, the more specific notion of 'repositories' commonly refers to digital storage areas for electronic publications or data collection that contain associated metadata. They can be designed either as large centralised databases or consist of distributed cross-linked systems. Metadata aggregators collect the metadata of various databases and provide a common point of entry and contact, while the actual data is stored on locally distributed systems organised differently (e.g. Europeana). Moreover, it is becoming increasingly understood that RIs also need to take a fundamental role in digitally preserving our research data and knowledge, a digital version of the long-term stewardship of libraries tracing back to the foundation of the Library of Alexandria in 300 BC. The societal responsibility to digitally preserve and provide access to our collective or individual cultural and social heritage often falls to public investment on behalf of society, which then enables researchers to exploit content that might otherwise be lost to posterity, and also to address societal challenges with both research and societal impact.

The significance of RIs remains beyond controversy amongst the scientific community and science funders and yet RIs still deal with considerable challenges. Often, infrastructures are planned on a long-term scale, but their financial support is only assured for a short period. Discontinuity in funding for established Research Infrastructures risks losing expertise and skilled personnel and degrades the innovative potential of both a region(s) and discipline, which equates to a waste of investments that have been made to date. This is why long-term availability and robustness of

RIs is essential for researchers to invest time in learning how to make best use of them. Often, sustainability cannot be guaranteed, which is problematic because sustainability is a central precondition for acceptance and trust within the scientific community. Data providers and data users need to rely on the fact that digital data can be identified in a unique and persistent way so that they can serve as scientific references over the long-term.

These challenges need to be addressed by the following strategies:

**Data centres**: In any field where there are no services provided, to safeguard and ensure access to research data, data centres need to be developed as soon as possible; decentralised, distributed models are more likely to fulfil the needs of researchers. They should observe key certification standards in order to provide trustworthy data:, e.g. ISO 16363 (Audit and certification of trustworthy digital repositories), DIN 31644 (Criteria for Trustworthy Digital Archives, or the Data Seal of Approval (datasealofapproval.org).

**Sustainability**: This strategy includes a combination of providing stable services based on proven procedures and technologies and the commitment by funders to invest in innovation where the cycles may be determined by a number of external factors or constraint. But whatever these cycles are, researchers need to be assured that infrastructures will be supported for the duration of their research and its long tail, which is recommended to be for one research generation (30 years). However it is recognised that funds are limited, and in order to optimise investment, prioritisation exercises should be conducted prior to establishment of new infrastructures, and an evaluation and exit strategy should be developed in consultation with the community. Where excellent infrastructures are already in broad usage, sustainability is critical.

**Trust**: RIs need to be reliable, persistent and trustworthy. Certificates like the Data Seal of Approval established by "Data Archiving and Networked Services (DANS)" can create confidence within the scientific community; the preferential funding of such infrastructures undergoing certification and respecting international standards could represent a means of an active regulation of this process. The provider can make sure that the citations referring to stored data remain identical on a long-term scale due to persistent identifiers. Implementing an agenda of trust, different players have to collaborate in order to increase the amount of certified e-infrastructures. In addition, when a project comes to an end, the transmission of the data to a certified e-infrastructure should be defined as a requirement by funding agencies. Finally, metrics permitting a reliable quality assessment of infrastructures have to be developed.

**Good governance**: an appropriate structural setting ensures outstanding and successful RIs that are tailored to the needs of their users; in this context, supporting institutions like the

Academies are asked to conduct regular evaluations and need-oriented surveys addressed to the interested communities in order to provide basic information for funding.

**Open access**: together with science funding institutions, the scientific community has to fight for free and non-restrictive access to scientific digital data as they are of crucial significance for the development of independent science and research.

**Standardisation**: the standardisation of file formats and data codes (e.g. PDF/A or TEI when it comes to text information), of metadata and data models (e.g. Dublin Core, RDF, OWL), of access protocols (RESTful, SPARQL-Endpoint etc.) or of archiving information systems (e.g. OAIS) has to be encouraged and needs to become a crucial issue in the agenda of the ALLEA Working Group and for further important coordinating players such as DARIAH in order to guarantee the interoperability between data resources.

**Clustering**: in accordance with recommendations made in the context of the EU research programme Horizon 2020, the linking and networking of data sets as well as the clustering of RIs have to be prioritised; this may significantly add value to the present-day situation where infrastructures are still isolated with regard to content; clustering is facilitated by organising data as linked open data in the semantic web or by establishing web services.

**Data Curation**: the operation of data and information repositories should not exclusively aim at long-term storage of information (archiving), but should embed "data curation" as well, that is to say a long-term disposability of information immediately accessible and provided in data formats that can be used directly and comprehensively (re-use and valorisation of existing data, data mining etc.)

**Recognition**: qualitatively outstanding RIs need well-trained staff that can bridge researchers' needs and IT-requirements; ALLEA supports specific trainee programmes for the Digital Humanities as well as an academic recognition of efforts with regard to RIs.

This group recommends the use of funding instruments and mechanisms that enable and facilitate the maintenance and development of durable infrastructures upon which researchers can build their research, experiments, analysis and tools. On the personnel side, we recommend adoption of metrics that supports data citation, publication of data, and due recognition for the skills and expertise required. Trusted digital archives and repositories should be prioritised by funding agencies and by researchers. Funding models that rely in part on leveraged funding and levels of cost recovery through paid-for services should be pursued where appropriate, but the core task of maintaining a long-term digital RI should primarily be supported by public investment for the purposes of research and innovation, education, and an enhanced, enriched multi-cultural society.

## Examples of data centres for humanities data:

Single-Site Digital Repositories

**Data Archiving and Network Services (DANS) (Netherlands)**
Established in 2005, DANS is the Dutch national archive of digital research data. Its mission is to promote sustained access to digital research data. It provides archiving services, training in reuse of data, and outreach. It is also the home of the Data Seal of Approval, an accreditation guideline for digital repositories.
http://www.dans.knaw.nl/nl

**UK Data Archive (UKDA)**
UKDA currently curates the UK's largest collection of digital data in the area of social sciences and humanities. It was founded in 1967; the first downloadable datasets became available in 2001. The Economic and Social Research Council (ESRC), JISC (formerly Joint Information Systems Committee) and the University of Essex primarily fund it. UKDA acts as a broker to the collections it is committed to curating and is a national project in the UK. The project also includes the UK Data Store, an online self-archiving system that collects, curates, and preserves a range of digital objects. The project provides documentation and workflows to users on how to access and share data in the system.
http://www.data-archive.ac.uk/

Multi-Site Digital Repositories

**Digital Repository of Ireland**
The Digital Repository of Ireland (DRI) is the national Trusted Digital Repository for social and cultural data (historical and contemporary) held by Irish institutions; it provides a central internet access point and interactive multimedia tools for use by the public, students and scholars. Developed as a government-funded project with six Irish academic partners, it is also supported by collaborations the National Library of Ireland, the National Archives of Ireland (NAI) and the Irish national broadcaster RTÉ.
http://www.dri.ie/

**Huma-Num (France)**
Established as a large-scale infrastructure (TGIR) in France, this multi-site institution aims to facilitate the digital turn in research in the humanities and social sciences. To carry out this mission, Huma-Num is built on an original organisation relying on a social structure (collective cooperation) and a technological structure (digital long-term services) at the European and national levels, based on a wide network of partners and operators. Its main mission is to offer a service matrix focused on a) long-term preservation b) tools and methods, and c) presentation and visualisation. It offers these services on a national level.
http://www.huma-num.fr/

Institutional repositories

**Geisteswissenschaftliches Asset Management System GAMS (Graz, Austria)**
Since 2003 the Centre for Information Modelling Austrian Centre for Digital Humanities of the University of Graz provides an infrastructure for a variety of Digital Humanities projects. The main aims of the

institutions are: a) applied research in the field of information processing in the humanities, b) support and cooperation for research projects in the humanities, and c) repository for long-term access.
http://gams.uni-graz.at/

**Data Center for Humanities an der Universität Köln (DCH) and Cologne Center for eHumanities (CCeH) (Germany)**
The DCH is a central facility of the Faculty of Humanities of the University of Cologne. Its mission covers a) permanent safeguarding, availability and presentation of digital research data, b) improving the visibility of active research projects and coordination between similar research projects, c) strengthening the existing interdisciplinary structures of the Faculty of Humanities, d) increasing the digital skills for postgraduates and staff of the university, e) support of on-going research projects both methodologically and technically, f) support of institutes and chairs in project development, acquiring external funds, and project implementation and g) networking on a national and international level with institutions in the field of e-Humanities.
http://www.cceh.uni-koeln.de/

# Digital Humanities Networks and Organisations

The recent years have seen the development of several European networks and organisations that contribute to better management of and accessibility to primary resources for researchers in the humanities. These endeavours cover various aspects of scholarly needs in the digital domain, and accordingly take different forms:

- Scholarly networks such as the European Association for Digital Humanities (EADH http://www.allc.org/), which helps Digital Humanities specialists exchange their expertise

- Research Infrastructures such as DARIAH and CLARIN that coordinate the stabilisation of services to support digitally-enabled research and teaching across the humanities and arts

- Large scale research networks such as the Research Data Alliance (RDA - www.rd-alliance.org), which promotes the international cooperation and infrastructure required for advanced data-driven innovation, by building the social and technical bridges that enable data sharing and exchange

- Library (LIBER) and archival networks (e.g. ENArC http://enarc.icar-us.eu/) that coordinate the representation of and access to digital assets across European cultural heritage institutions

- European portals (Europeana) making available to a wider public the descriptions of the content of cultural heritage institutions

- And of course ALLEA, which established the E-Humanities Working Group, and facilitates research into digital requirements through efforts such as the SASSH survey.

Although a good and natural link exists between scholarly communities and research infrastructures, there is still a need to clarify and communicate the role of the various actors and define their respective roles with regards to research. In particular, the various cultural heritage and infrastructural networks should facilitate access to primary material in both physical and digital forms.

## SASSH

In September 2013, the Union of the German Academies of Sciences in close cooperation with the federation of All European Academies (ALLEA) launched a "Survey and Analysis of Basic Social Science and Humanities Research at the Science Academies and Related Research Organisations of Europe" (SASSH). The primary objectives of the SASSH initiative were to compile the first ever inventory of basic SSH research in the social sciences and humanities (SSH) undertaken at and/or by these organisations and to create transparency in the relatively little known academies' research landscape. The secondary and long-term objective was to identify similarities within the academy research landscape, and therewith the potential to connect the research projects of the European science academies to form pan-European research clusters on matters of European cultural heritage and identity.

Enhancing the coherence and connectivity between projects across Europe requires digital resources and tools accessible to and useable by all. These ease collaborative research processes while also fostering innovation. A European research programme must have at its disposal common, compatible systems for accessing, collecting, generating, sharing, analysing, storing and disseminating data and results. Where such systems are already in place, their compatibility should be harmonised. Where such systems are lacking, they should be implemented. The SASSH initiative thus placed a special focus on digital practices in academy-based SSH research. The survey investigated to what extent and purposes scholars use digital tools in their research, needs and wishes for future research tools, the popularity of electronic open access publishing and data archiving, and the existence of data standards and support and training for Digital Humanities methods at the institutional level.

The survey reached over 600 SSH research projects Europe-wide and will conclude with a comprehensive publication of its findings in the spring of 2015.

## Recommendations for Networks

- Foster a higher technical coordination between European networks in the humanities. In particular, issues related to hosting, basic standards or licensing principles should be dealt with in a uniform way and the corresponding services made better available to scholars. See for instance the Marie Curie network DiXiT, which provides a European training framework in the domain of digital editions — (http://dixit.uni-koeln.de/)

- Encourage research infrastructures to cover more disciplines in the humanities in a coherent way, in order to ensure the widest sharing of experience in the domain of digitally-based research methods

- Give recommendations and guidance concerning the development of local partnerships between researchers and cultural heritage institutions (standard cooperation agreements, license frameworks and technical workflows)

- Support the full representation of humanities data in research data sharing networks and platforms that historically focus on science or social science data

- Encourage the adoption of outputs by working groups at the Research Data Alliance

- Encourage all Digital Humanities funded projects to establish a sustainability plan in consultation with the DARIAH research infrastructure.

# Dissemination and Access to Scholarly Information

The whole idea of scholarship is oriented towards maximising the dissemination of research results. Carrying out a research activity is indeed about exploring territories, where knowing what others are doing, what their most recent advances are and what projects are being undertaken is essential to make sure that one's own research actually goes beyond the state of the art and can be situated within a larger corpus of discoveries. Communicating results is thus an essential activity in one's academic life, all the more so because the assessment of such communications through peer review mechanisms impact the capacity to get institutional recognition and thus the financial means to carry out further research.

Until now, books, journals and to a lesser extent conferences have been the main media for the publication of scholarly results. Still, the recent period has shown the limits imposed by such frameworks because of the incredible increase in publication prices, the reduction of actual services provided by publishers, and also the reduced visibility offered by traditional publication means, in contrast with what the internet technologies should be able to offer.

In this context, scholars increasingly demand ways to quickly and cheaply disseminate their research results while keeping the benefits of the assessment of their works by their peers. Higher education and research institutions should thus take action to favour publication platforms that allow an immediate online availability of results, at a price that is affordable for them.

It is thus essential to provide guidance to scholars as to how they can make their research results widely accessible according to open access principles. From a technical point of view, there is a need to identify a strong network of publication repositories where scholars could deposit their research articles in open access form in parallel with traditional publication in conferences and journals. They should also be provided with publication platforms for the management of open access journals. Finally, adequate open repository infrastructures should accompany the need to disseminate openly digital scholarly sources and data sets. DARIAH offers services on all these aspects through the OpenEdition journal platform, and also through its awareness activities around open access principles within the arts and humanities.

## Recommendations on Dissemination

- Help research and higher education institutions define an informed open access policy based on the interest of public research
- Encourage research and higher education institutions to define assessment profiles that take into account a variety of publication forms
- Widely inform scholars about the potential of new publication forms (e.g. blogs) and dissemination channels (e.g. publication archives)
- Support the necessary public technical infrastructures for open access dissemination of research results.

# Publishing and Licensing Data

The emergence of Digital Humanities scholarship and the development of digital research methodologies are contributing towards a revolution in traditional scholarly publishing and the system of citations built upon it. This revolution is partially based on the innovative interdisciplinary, virtual nature of the Digital Humanities, but is also linked to the broader publishing context which is challenged by new technological platforms, data formats and consumer expectations. These challenges, coupled with the relative ease by which large datasets can be compiled and published online by a researcher without regard for best practices, mean that it is necessary and timely to consider the development of a strategic framework for the publication of Digital Humanities research.

In Europe, policy makers and the stakeholder community (for example the European Commission, the Open Knowledge Foundation) are advocating the publication of openly licensed research data and research publications in the Sciences and the Humanities. The application of open licences such as Creative Commons offers benefits such as the efficient dissemination of results, reusability of research, knowledge transfer between disciplines, and contributing to the public good. To facilitate the dissemination of openly licensed research, a number of Open Access publishing models have been developed, including Gold, Green, Platinum and Hybrid models.

## Open access distribution models:

**Traditional**

Subscribers pay to have unrestricted access to content

**Gold**

Author pays APC (author processing charge) substituting for subscription fees – article is immediately available to all

**Platinum**

Publisher obtains grant, sponsor, or donor to cover cost, with little or no author fee

**Green**

Authors deposit their pre-publication manuscript in an open repository

**Hybrid**

Combination of traditional and gold access models

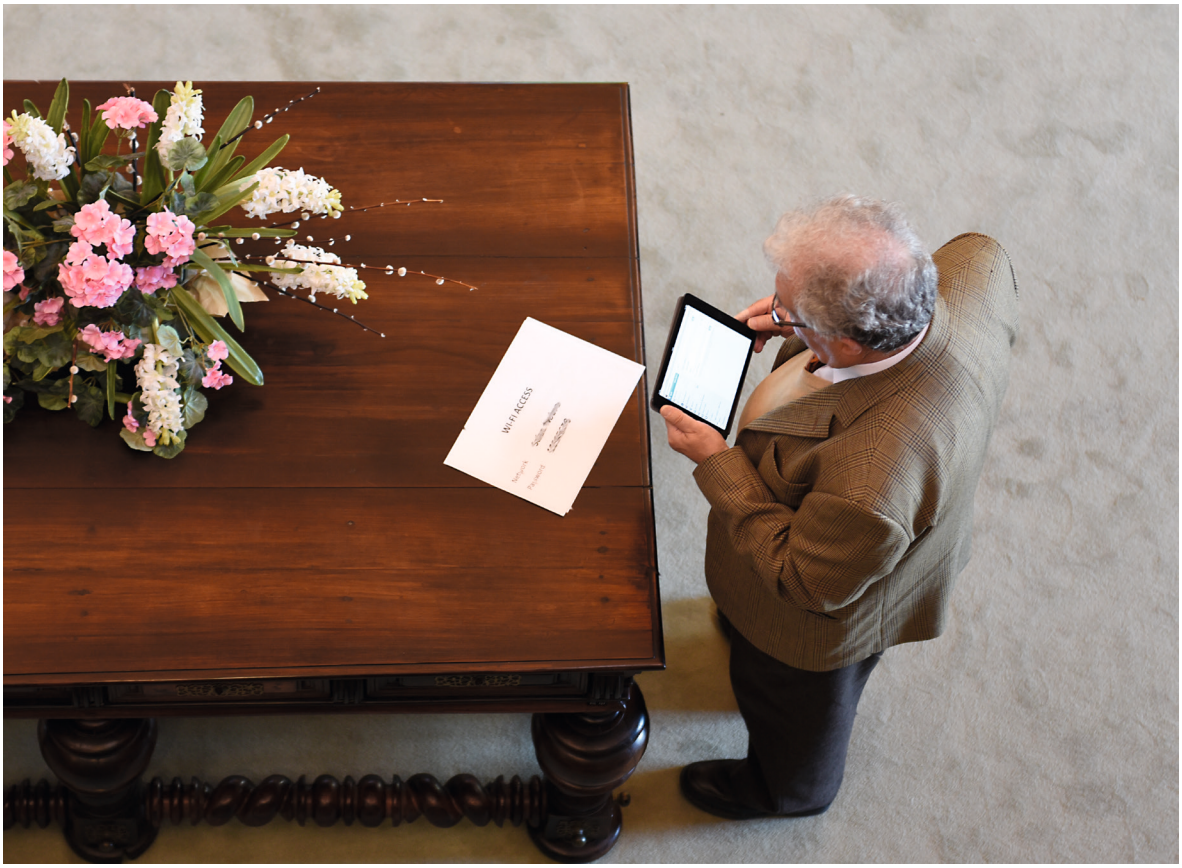
**Delayed**

Publisher provides open access for no additional fee after an embargo period of 6-12 months

Despite the tangible benefits to Open Access publication, Digital Humanities scholars must be mindful of intrinsic issues of copyright and data protection which may be associated with their research. National and European legislation is designed to protect intellectual property and individual privacy and these aims must be balanced with the desire to disseminate research openly.

The E-Humanities WG thus recommends:

- Develop strategies and explore best practices in order to develop a Digital Humanities publishing framework
- Support training in Digital Literacy for Humanities scholars
- Refine citation matrixes and develop scholarly metrics for Digital Humanities data
- Create awareness of relevant tools and infrastructures for the Digital Humanities
- Support open licensing for reuse and incentivise reuse actively
- Contribute to the development of Standards, e.g. Citation
- Explore best practices of editorial integrity within the open science paradigm
- Contribute to the development of legal solutions for publishing in global networks
- Support sharing and actively further the open science paradigm
- Support sustainable infrastructures for digital publishing



# Long-term Digital Preservation

Digital Humanities projects produce a lot of content, from digitised artefacts through to searchable databases, from interactive websites to large-scale 3D platforms. In current practice, the intended destination for these projects is often online via a project website, or on a personal/departmental research page on a university server. While this method of storage is useful for short term access and dissemination, it does not meet the requirements for long-term digital preservation on at least two counts. First, website storage does not fulfil the technical requirements for trusted digital preservation, according to internationally accepted certifications of trustworthiness mentioned above (e.g. TRAC, DSA, ISO 16363). Second, website links are easily broken, lost, or abandoned when the project comes to an end or the researcher changes positions or institutions. For long-term access to be assured, DH projects need to be archived and preserved in managed ways.

In addition to completed DH projects, scholars produce a significant amount of data during their research, and while the resulting project from that data may be shared and disseminated through trusted channels, the research data itself is often stored on the researcher's computer or external storage media, and therefore faces the same risks of loss, degradation or inaccessibility as any other data stored in this way.

Long-term digital preservation requires attention to the life-cycle of digital data; it requires ongoing management to ensure that file formats continue to be accessible, and that the data is protected from media failure, corruption and abandonment. Depositing Digital Humanities data in a trusted repository is one way to extend the lifespan of Digital Humanities research, and to ensure it is accessible in the future. Initiatives to archive and preserve born-digital material, including web archives and the archiving of social media, are also important in order to capture the breadth of material being generated as part of our cultural record. Accordingly, we recommend that best practice in Digital Humanities research includes planning for the long-term preservation of research data and research outputs, as well as the preservation of required software and tools, and project communications. Planning for preservation should be conducted at the beginning of a project, and where possible, digital preservation should be itemised as a cost in research grant proposals.

# Policy

The vision of an open society of science and scholarship in Europe and beyond depends mainly on the possibility of sharing data and knowledge. Sharing data requires a free and open access to research data for everyone – a widely accepted precondition.

> '*How policy makers and funders can target their limited resources at so many points of the data sharing ecosystem for maximum social and economic benefit is an enormous question to which there are no simple answers. But two things are clear: that investment at all these points is necessary to create a fully realised data sharing system; and that gaps and redundancies in investment can best be avoided by a coordinated approach on the part of all agencies – governmental and non-governmental – that make research policy and fund research activities.*' (Research Data e-Infrastructures: Framework for Action in H2020).

Given the constant growth of data produced in the humanities and social sciences, funding policies are needed to foster a secure, sustainable and trustworthy open access to the results of publicly financed research. In 2012 the European Commission recommended to the Member States that they define clear policies for open access to scientific information, including publications and primary research data. The results of research should be accessible to the public for free in order to ensure the use and re-use of data. The responsibility to implement adequate policies implies different levels such as transnational, national or local institutions and organisations (EU, ALLEA, national ministries of education/universities/science, funding institutions, foundations etc.).

In the tradition of activities for promoting the 'Public Understanding of Science' in the UK, the Research Councils have already established a whole set of policies on topics like open access. The policies are related to funding mechanisms helping to implement them. Annual impact reports provide information on the progress achieved. In Germany, the Alliance of Science Organisations has adopted principles for the handling of Research Data including the topics of long-term preservation and accessibility.

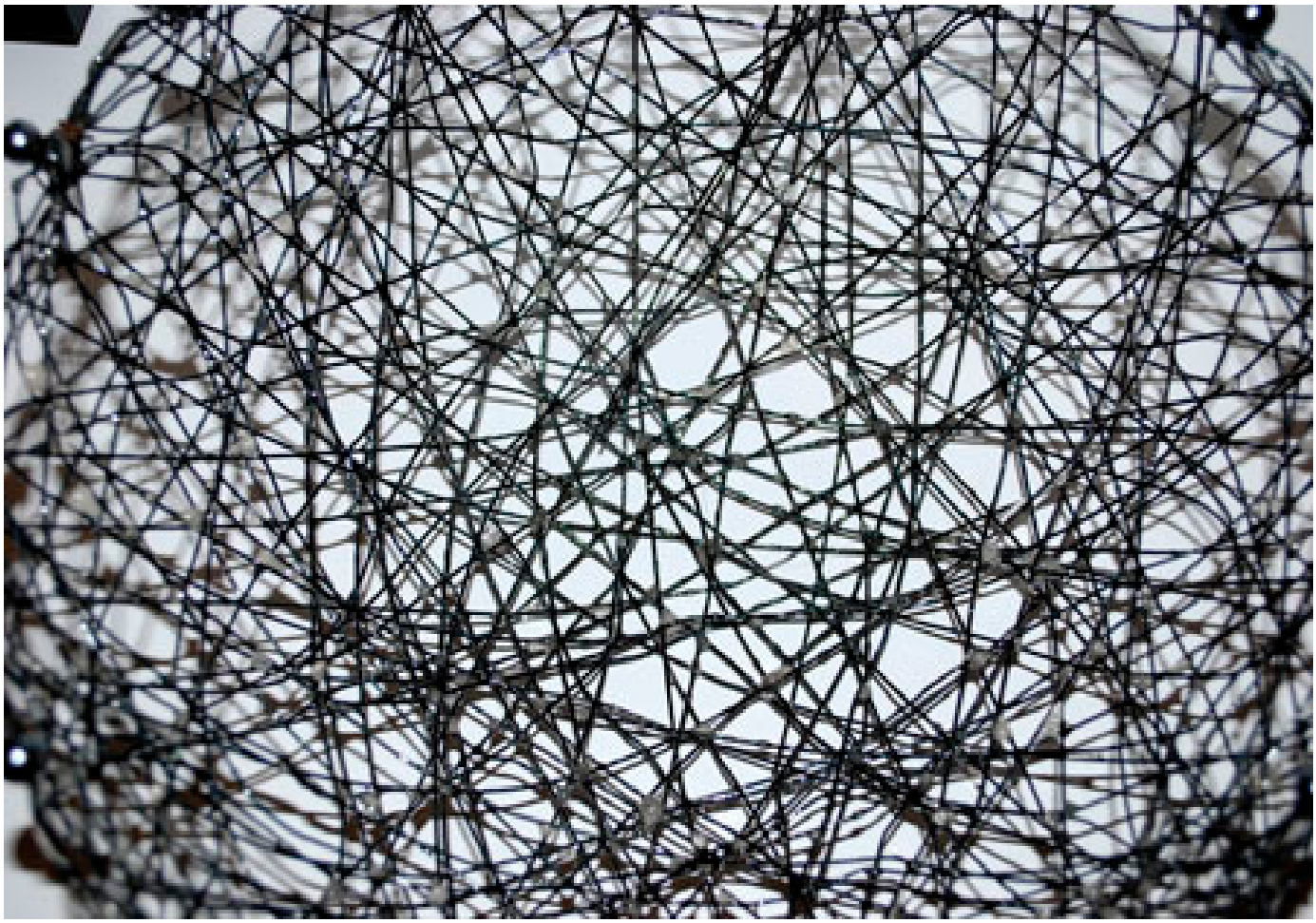
## Research Data

Although several nations have already established policies on data management and especially on the free access of publicly funded research results (data and publications), the European research area at large is still encountering major problems like data loss or restricted availability to research findings.

**Establishing policies:** First of all, policies on data management and open access should be introduced on a national level where they are still lacking, in accordance to supra-national initiatives. The implementation of guidelines or policies is usually preceded by substantial debates in order to create common grounds between all stakeholders involved in the scientific process on the main principles of dealing with research data.

**Embedded policies:** Policies on research data on a national level should be backed up by corresponding regulations of the funding organisations in order to improve the impact on the long-term preservation and availability of publications and data.

**Evaluation of policies:** The impact of policies for research data should be checked regularly, for example with annual progress reports.



## Recommendations for Research Policy

- Open access by default: One of the main objectives of a policy should concern free and open access to research results that are publicly funded.

- Access limitations: Although free access as a basic guideline is strongly recommended, reasonable interests of data producers (intellectual property right) and mediators (embargos) should be treated in accordance with internationally accepted licences or standards like Creative Commons; embargos for the free dissemination of printed articles depend on disciplinary conventions but should not excess the period of 12 months.

- Open data: Policies on open data principles should always include both primary research data (with associated metadata) and publications.

- Data management plans: In order to guarantee basic standards of data management including their long-term preservation and availability, DMPs can be defined as a pre-condition for the eligibility of a research project for public funding. Depending on locally existing facilities, the whole life cycle of research data including the deposit in a data archive can be framed by funding policies.

- Freedom of research: A strong opposition against open data is formulated by insisting on the freedom of research. The argument here is that no one should be forced to share his or her data and results in repositories or other institutions they are not interested in. It is of great importance to balance between individual interest and the common good, because otherwise many scholars will not accept a policy of open data.

# Training for Digital Humanities

Scholars whose work lies firmly in the domain of Digital Humanities should be offered training in the skills necessary to exploit the technical potential of the platforms and applications relevant to their work. This may seem like an obvious statement, but many DH scholars are self-taught when it comes to the digital aspects of their work, and would benefit from more thorough training. Importantly, training for DH scholars needs to be specialised and targeted at Digital Humanities research problems and requirements, and not just vaguely aimed at 'digital literacy'.

Computational approaches to humanities research are increasingly being incorporated into contemporary research projects, and similarly an understanding of humanities and social science perspectives is increasingly required for successful ICT projects (for example, witness the focus on the integration of SSH research into ICT programme calls across the EU's Horizon 2020 funding scheme). Academia and industry are also widely demanding or expecting higher level digital skills from scholars and employees.

In order to respond to the growing integration of digital practices and approaches in humanities research, students and researchers should be offered training in digital skills that is commensurate with contemporaneous educational expectations and prepares them to understand and incorporate digital best practices in their work. Early training in data management and digital archiving, for example, contributes to the longer-term sustainability of Digital Humanities research by ensuring that good data practices shape the approach to a research project from the beginning. Early training is more cost effective than the effort required to reconstruct or recover data sets that were not treated in a sustainable fashion from the outset. This training can be targeted in a number of ways:

- Exposure to the unique methodologies of Digital Humanities could be included in humanities programmes, and similarly, hands-on digital training could be incorporated directly into Humanities curricula at the third level so that the potential for exploiting digital tools is revealed before a research programme is fully conceived. Considering digital approaches to research problems has the potential to shape research methodologies, whether the project has computational aspects at its core or as one aspect. Funding for training should also be recognised as a suitable cost in research grant proposals.

- Inversely, research questions and theoretical frameworks associated with humanities research could be part of the training curricula for software engineers whose work will directly impact everyday questions that occupy humanities researchers.

- Higher Education courses are required to train the cadre of data scientists and digital archivists required to facilitate work in the Digital Humanities, and to maintain research infrastructures. These courses, with suitable accreditation, should target the areas of data science, data analytics, digital archiving, digital libraries and repositories, digital preservation, data management, digital curation, knowledge organisation, and information systems. Appropriate training provides the basis for timely career progression.

# Recognition and Career Progression for Digital Humanities

Early career researchers often collect large amounts of structured data. A typical PhD in the humanities often involves structured analysis of some set of primary data, for statistical analysis or just for ease of reference for the researcher, as need may be. All too often such data exists in spreadsheets or simple databases on personal computers, at some point becoming obsolete and no longer readable. One reason for this is the lack of recognition of the work that goes into proper digital curation. Young researchers react by giving priority to the kind of activity which will give them jobs, namely research publications. This bias towards publications may lead to a certain short-sightedness: it may be efficient in the short run not to take the extra care that is needed to prepare your data for use by others, but in the long run this leads to wasted efforts.

To remedy this it is essential that digital curation and the creation of data collections that are useful to outsiders must be given proper recognition in all processes where researchers (especially young ones) are evaluated for employment, tenure or project grants. Project proposals are evaluated for their dissemination plans, not only in terms of planned publications but also data dissemination. Many institutions provide templates for evaluation: these should always include such considerations. Research progresses through a collective enterprise, and a well organised and easily available research infrastructure may be as important as a brilliant theoretical contribution.

It is therefore necessary to make sure that there are jobs for digital humanists at all levels - PhD scholarships and postdoctoral positions, but also mid-level jobs and full professorships – and that candidates for these jobs are evaluated for their merits in the Digital Humanities. In many universities, Professorships in Digital Humanities have already been established or are in being planned with a particular focus. In Basel, for example, the profile of the Chair emphasises a) securing long-term access to digital research information, b) preserving access to digital sources and providing for reliable citation and c) preserving cultural heritage through digital procedures, whereas in Bern, particular attention was given to methodological and technical competences like computer-based analysis of texts, visualization of complex data, digital editions or digital valorisation of archived records. Also, it is important that there is recognition and clear career progression opportunities for archivists and other technical staff who are not on the academic track, but are indispensable for Digital Humanities projects. To maintain and develop the Digital Humanities in academic institutions, it is important to offer this staff – often attracted to work outside of academia – with real opportunities.

# Conclusions and Recommendations

Europe has a vibrant, rich cultural heritage and a tradition of excellence in humanities scholarship. Digital Humanities is an emerging discipline that includes a broad range of activities in research, teaching, and cultural production, and uses digital technologies to preserve cultural objects, to make them more accessible to researchers and the public, and to analyse and mine their information in new and previously impossible ways.

Digital Humanities exists in a complex ecosystem of funding constraints, institutional career progression models, changing publishing models, technological developments and an increasingly networked world. How can we ensure Europe retains a leading position in the Digital Humanities, and what actions can European research authorities and actors take to enable leadership and excellence?

## Recommendations

**Take a long-term view:**
Sustaining long-term archives of unique and important cultural artefacts is critical for Europe's leadership in Digital Humanities. We recommend a move to funding models which are not project-based, the certification of digital infrastructures, appropriate funding evaluation, pan-European policies and strategies, and adoption of best practice in digitisation and metadata standards and vocabularies. Researchers should look to large organisations such as the Research Data Alliance and DARIAH for best practices.

**Encourage openness:**
An open approach to data enables research integrity, increased secondary research and cost-effective data production. Open Access should be incentivised and increasingly mandated, data management plans should be required with all funding proposals, and data archiving costs should be included as eligible costs. Training and open repository services should be openly available, and standardised data citation should be adopted and recognised.

**Support your people:**
In Digital Humanities the people are no less important than the infrastructures and technologies. Career progression models should include recognition for the importance of data activities including data design, collection, curation, and management. Specialist training should be funded, openly accessible and certified. The roles in data management such as data librarian, data scientist, data archivist, should be recognised in the research community as trained and skilled roles.

# Appendix: Resources and Further Information

### Digital Humanities Readers:

S. Schreibman/R. Siemens, & Unsworth, J. (Eds.), A Companion to Digital Humanities, Oxford, 2004.

Melissa Terras/Juliane Nyham/Edward Vanhoutte (Ed.), Defining Digital Humanities. A reader, Surrey: Ashgate 2013.

Jerome McGann, A New Republic of Letters. Memory and Scholarship in the Age of Digital Reproduction, Harvard UP 2014.

### Reports:

European Strategy Forum on Research Infrastructures (ESFRI) (2011), Strategy Report on Research Infrastructures. Roadmap 2010. Luxembourg: Publications Office of the European Union. DOI: 10.2777/23127
Download: http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf#view=fit&pagemode=none

Makarow, M., Žic Fuchs, M. & Moulin, C. (2011), Research Infrastructures in the Digital Humanities. Science Policy Briefing 42, Strasbourg: European Science Foundation
Download:http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/be_user/research_areas/HUM/Strategic_activities/RIs_in_the_Humanities/SPB42_44p-5oct_FINAL.pdf&t=1395833830&hash=59c480308f03879a091a5880858ca7dbd22667ab

O'Carroll, A., Collins, S., Gallagher, D., Tang, J., & Webb, S. (2013), Caring for Digital Content, Mapping International Approaches. Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy. DOI: 10.3318/DRI.2013.1
Download: http://dri.ie/caring-for-digital-content-2013.pdf

### Regulations and certifications:

Data Seal of Approval: http://www.datasealofapproval.org

Trusted Digital Repository: http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

DIN 31644: Information and documentation – Criteria for trustworthy digital archives:
http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&artid=147058907&languageid=en&bcrumblevel=3

ISO 16363:2012: Audit and certification of trustworthy digital repositories: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

Standards of trust of the UK Data Archive:
http://www.data-archive.ac.uk/curate/trusted-digital-repositories/standards-of-trust

Nestor Seal for Trustworthy Digital Archives: http://www.langzeitarchivierung.de/Subsites/nestor/DE/nestor-Siegel/siegel.html;jsessionid=887FED409200A3F1D75E2EA42EF5F775.prod-worker2

## Digital Resources:

- Digital Research Tools: http://dirtdirectory.org/
- DH-Tools: https://de.dariah.eu/dh-tools
- Digital Toolbox: http://www.infoclio.ch/de/node/130300
- Digital Curation: http://www.dcc.ac.uk
- Data Repositories: http://www.opendoar.org http://service.re3data.org
- DARIAH e-infrastructure: http://dariah.eu/activities/e-infrastructure.html

ALLEA

ALL European Academies

c/o Berlin-Brandenburg Academy
of Sciences and Humanities

Jaegerstr. 22/23
10117 Berlin
Germany
tel +49 (0)30-3259873-72
fax +49 (0)30-3259873-73

secretariat@allea.org
www.allea.org
twitter: @ALLEA_academies



**ISBN 978-3-00-049483-3**