Mapping the Landscape of eResearch
Text – Image – Annotation

International Workshop, Berlin 2012

Introduction, Scientific Implications

Jan Simane

In 2009, when four Max Planck libraries of the Human and Social Sciences Section, together with the Max-Planck Digital Library, decided to cooperate in developing an environment for access to and research on their digital collections, the discussions were determined by two principle interests. The first was the integration of already existing and running projects and initiatives as well as the corresponding knowledge, experience and expertise into one coordinated and cooperatively organized infrastructure in order to avoid redundancy and boost synergy. The second was the driving curiosity and desire to develop new applications and tools for analyzing, encoding, annotating and editing first and foremost digitized texts within the meaning of a virtual research environment. Despite the clearly expressed accordant goal of all the partners the project is determined by one important characteristic: regard for the specific methods, source types and research tools of each represented discipline and, consequently, the requirement to satisfy the expectations of the scientists concerned. Let me go over the three mentioned topics – the situation before the project's departure, the challenge to develop the appropriate tools and formats and, finally, the ambition to allow qualified users an active role in working within the application's infrastructure – and carve out the contour of the project's concept step-by-step.

One typical feature of the Max Planck Society is the permanent swinging between autonomy and individualism of every single institute on the one hand, and the endeavour to centralize and unify administrative and operational infrastructures as far as possible on the other. One outcome of the latter was the foundation of the Max-Planck Digital Library in 2006, which is anything but a library. Rather, it is a central institution for IT-services, the supply of digital resources and the development of new IT solutions. In previous years the situation was rather heterogeneous. When we focus on the humanities and on corresponding digitization projects we quickly come to two important stakeholders, the Max-Planck-Institute for the History of Science here in Berlin and the Max-Planck-Institute for European Legal History in Frankfurt. The Berlin Institute for the History of Science offers on its website an extensive collection of digitized source books, websites of research projects, digitized research data, the Institute's publications as well as online exhibitions and databases. The Institute has a long experience in digitization processes, transcription routines, and was a leading partner in the development of web-based applications like *digilib* or portals to digital collections like the Archimedes Project and Echo. Undoubtedly, the MPI for the History of Science was and is one of the most active and successful institutes in promoting the digital medium as an integral element of a complex research infrastructure and its experience and expertise was always an important guidepost for other institutes with similar intentions. And although the Berlin colleagues are not immediate project partners in the current initiative we are presenting here in this workshop, we are very grateful for their willingness to exchange ideas and expertise for the benefit of this matter. The MPI for European Legal History, on the contrary, is a digitization expert more from the librarian point of view. It offers an impressively extensive collection of digitized legal source texts from the 16th to the 19th centuries which has been compiled in several projects of recent years and focuses on the traditional role of the library as the most important research tool in the humanities. Of course the collection goes far beyond a mere aggregation of digital copies of the printed originals. The intention is rather to harness the potential of the digital technology and offer varied forms of connecting information from shared sources and of different types. As a consequence of the previously mentioned heterogeneity of the Max-Planck Society the plurality of methods and concepts not only in the research field but also concerning digitization is an uncontested matter of fact. Therefore the requirements on and experiences with digitization campaigns, for example in the Bibliotheca Hertziana in Rome, are significantly different to those of the MPI for Legal History. For decades one of the core areas of research in the Roman MPI for Art History has been the extremely complex topography of the city of Rome. The corresponding library holdings, like historical city guides, have been digitized to a large extent and incorporated into a network of digital collections like maps of Rome, architectural drawings or a web-based virtual repository of research data from various projects of the institute. Here, the linking of single details of both texts and images to related data in other contexts for example, or the use of illustrated books as a source for images were fundamental requirements and decisive for the definition of data formats and digitization standards.

These are only a few examples – and I hope I may be forgiven for omitting others that are no less important – which illustrate the situation in the Max-Planck-Society before the project started: a keen interest in the potential of digital technology, year-long experience in digitizing, cataloguing, encoding and providing various types of sources, single solutions for particular requirements and different forms of infrastructure. But also a lot of overlaps and common interests. The awareness of this fact and the willingness to cooperate for the purpose of pooling know-how and investments were the initial motives for developing a common project. But this was just one side of the coin. The other is more challenging and focuses on the basic issue of the nature of the digital object, its identity as a format in its own right and the related consequences for research methods and innovative tools. This is not the right place to delve into the ocean of discussions on the theory of computing in general and e or digital humanities in particular, not least because the project's goal is the development of an infrastructure for the practice and not a theoretical approach. However, a few theoretical considerations should be mentioned, also to justify naming the project "Digitization Lifecycle" which describes its identity of being a complex process and not a mere change of the aggregation state of one and the same object. As long ago as 2001, when the digitization of library holdings and similar material became a mass movement in many countries, strongly supported by related funding programmes, one of the leading experts in the field, Manfred Thaller from the University of Cologne, emphasized that a *digital* library should be more than a *digitized* one. And this is – in simple words – the resume of the whole background of our realization that a digital object is something else and significantly more than a mere copy of the analogue original. What exactly is this 'more' and what does it mean for research matters? Willard McCarty, who dedicated a good part of his professional life to this question, argues in a similar manner to Thaller: "the central function of computing for scholarly analysis is not building digital replicas of books, or what I call 'knowledge jukeboxes', but *modelling*".[1] Indeed, 'modelling' is the key term, rather key *concept*, related to the development of digital research environments. It is the basic demand to translate discipline specific methods, systems and conventions of analyzing and valuating subject-matters and processes into theoretical and abstract relation networks. In a second step these modelled concepts have to be formalized to make them compatible with the logic of computers. For information scientists the process of modelling and formalization are a matter of course, no doubt. But our project is based on the permanent

[1] Willard McCarty, Beyond the word: modelling literary context, p 11 (http://www.mccarty.org.uk/essays/McCarty,%20Beyond%20chronology%20and%20profession.pdf)

dialogue between the direction-giving group of librarians, who are also scholars and experts in their disciplines, on the one hand, and computer specialists on the other. Both parties have an influence on each other and we know that the humanities in particular have been remarkably manipulated by computer sciences over the last 25 years. As Emanuele Salerno from the Institute of Information Science and Technologies in Pisa pointed out, modern research methods in the humanities differ from traditional ones in the way they are forced to formalize their strategies and solutions. "If there has been a change in human sciences, it is not to be ascribed to the way in which they have used computers, but in their interaction with informatics as a science, from which they have drawn principles and operational methods."[2] Salerno emphasizes first of all new quantitative approaches to humanistic disciplines and the related qualities of the acceleration of procedures and retrieval as well as control over unprecedented quantities of data. These remarks were made ten years ago. Today, we can confirm that Salerno's assumption has become an uncontested reality in many branches of the humanities. But apart from the new potential of data management it should be recalled that digital objects, mass products of digitization campaigns, have a specific nature which is totally different to the digitized original: they are editable, they can be interactive, they are open to a broad public and they are distributed concerning hosting and storage. In other words, the original source, as already expressed by McCarty, undergoes a metamorphosis to a manipulatable existence whereby new ways to new knowledge will be paved. We could show a long list of dozens of web-based applications for humanistic studies all over the world where all or some of these principles have been effectuated. And our lifecycle project, which currently concentrates strongly on digitized texts, pursues the same goal, when an infrastructure for content enrichment, search and retrieval tools, cross-linking interfaces, annotations and encoding standards as well as various viewing options is being developed. Of course, we are fully aware that this is neither the first nor a remarkably original attempt in infrastructure building and our intention is to be as compatible and connectible to other initiatives as possible. Furthermore, we would like to learn from other projects, whether they are already finished or under construction, which is why we have organized this workshop and why we have invited all the experts. But we don't only want to discuss technical details and practical solutions. So we should also focus on the key question of whether our target group, the scholars and researchers,

[2] Emanuele Salerno, How computers affected the humanities, 2002, p.9. (http://jcom.sissa.it/archive/01/03/A010301/jcom0103(2002)A01.pdf.),

really need and require such tools or not. And this is the third and last topic of those initially mentioned.

For a couple of years one topic has been discussed extensively in the context of digital humanities, which can be summarized with the term the ‘IT competence’ of scholars and researchers. When in the early 21st century more and more digitized collections of texts and data were put online there was a certain unease regarding the gap for example between historians who used computers in a limited and rather traditional manner, and the potential that information science has established for processing and analyzing digital sources like archival material for instance. The discipline itself and computing remained two separate fields. The consequence was by no means the requirement to turn scholars into information scientists, but rather to develop tools and applications for the benefit of the research routines within the ‘information technology framework’.[3] This sounds easier than it is. In any case a new approach towards elaborated communication between scholars on one hand and IT experts on the other has been recognized as necessary, even more, the birth of a hybrid creature named ‘technician’ seemed to be promising. Someone who is also able to formulate discipline specific requirements and their possible technical solutions – at least in theory. In the meantime a lot has changed. Today, the rather hesitant use of computers and digital tools for research matters is certainly not a common phenomenon. And the IT competence of the so-called ‘scholars’ – to distinguish them from information scientists – has achieved an impressive level. However, can we really be sure that our projects and developments are based on this fundamental diffusion of the two spheres, the discipline-driven issue and the principles of information science? Or should we, on the contrary, confirm the polemically expressed suspicion of Hans Magnus Enzensberger, who, in 2000, wrote in his essay ’Das digitale Evangelium’: “Neue Medien sind immer auf der Suche nach unbekannten Bedürfnissen. An ihren Pionieren fällt eine kuriose Autonomie auf. Wenn sich Bastler, Ingenieure, Programmierer etwas ausdenken, sind sie ausschließlich an den Eigenschaften ihrer Spielzeuge interessiert. Der mögliche Benutzer ist für sie nur ein störender Ignorant.“[4]? Of course, it won’t be that bad. But we should acquire more certainty in this field than we have done at least in our lifecycle project. Also, if not primarily in this context, we are curious to learn from other projects and experiences.

[3] See Onno Boonstra, Leen Breure and Peter Doorn, Past, Present and Future of Historical Information Science, in: Historical Social Research / Historische Sozialforschung, Vol. 29, No. 2 (108), 2004, p. 90f.

[4] Hans Magnus Enzensberger, Das digitale Evangelium, Der Spiegel 2/2000, P. 94.

To conclude: our project started with a broad spectrum of experiences and expertise in digitization projects. It is based on the conviction that a number of new useful tools and applications can improve and facilitate the work with digital objects. And we do all this for the benefit of scholars and researchers. Right now we are in the middle of the project’s process, the perfect time to evaluate the first results and further steps.