Research partnerships, user participation, extended outreach – some of ETH Library's steps beyond digitisation

Michael Gasser

Research partnerships, user participation, extended outreach – some of ETH Library's steps beyond digitisation

Michael Gasser, Eidgenössische Technische Hochschule Zürich, ETH Library (ETH Library, ETH Zürich)

Introduction

Mass digitisation and the freely accessible online presentation of publications, archival material, photographs, images and objects with public domain or creative commons licences have become ongoing tasks for commemorative institutions. These days, numerous large-scale GLAM institutions, but also many smaller ones run digitisation centres and a steadily growing number of platforms to render their digital contents globally accessible for research and the interested public. ETH Library, ETH Zurich's main library, is a driver and part of this digital transformation. With its shift in strategic focus towards the "digital library" shortly after the turn of the millennium, on the one hand it geared itself towards providing licensed digital information resources within the university. On the other hand, the strategy also included (retro) digitisation from the outset.[1] A highly efficient DigiCentre and – partly in cooperation with other Swiss libraries – presentation platforms for digitised journals (e-periodica.ch), image materials (E-Pics), old books (e-rara.ch) or archival material (e-manuscripta.ch) were established.

The major, mounting interest in these online services is evident in their ever-increasing usage figures.[2] Consequently, the *Strategy 2015–2020* approved by ETH Zurich's Executive Committee for the university's collections and archives includes pressing ahead intensively with indexing and digitisation campaigns, and expanding them to include objects of scientific collections.[3] Although these collections are curated directly by biologists, entomologists, mycologists or other experts in the respective departments, ETH Library plays a central coordinating role in the evaluation and implementation of a digital infrastructure for them.[4]

The extension of digitisation to other collections and object types is merely part of the ongoing developments. In terms of the online services already established, other questions are

[1] Mumenthaler and Voegeli (2005) 67

[2] ETH-Bibliothek (2016) 70–71

[3] ETH Zurich (2014)

[4] ETH Library (2017a)

also becoming increasingly more central: how can the abundance of great potential and synergies that already exist in the digital copies already available and the established online platforms be exploited further for research and the interested public? How can the use and re-use of digital content be encouraged even further? How can users be involved directly in enhancing content? How can new user groups be addressed and reached? Which new and innovative services can be developed based on existing digital copies and established processes and platforms?

Naturally, there are no clear and simple answers to these questions. The dependence on key influential factors such as the content and quantity of the existing subject matter, available functionalities of platforms or the personal and financial resources on hand is too great. In this sense, the ETH Library initiatives and projects outlined below are merely intended as examples. The three topic areas into which these examples can be divided, however, reveal the general directions in which commemorative institutions can realise potential and expand services that go beyond pure digitisation: intensified partnerships with researchers, direct user participation and outreach intensification.

Research partnerships

Currently, of all the research branches, the *digital humanities* particularly offer excellent opportunities to generate added value from digitised content in collaboration between academia and commemorative institutions. ETH Library is heading in this direction on a cooperative project with the University of Zurich's Institute of Computational Linguistics, for instance. The basis for the project is the millions of retrodigitised and OCR-recognised pages of journals that ETH Library provides via e-periodica.ch, its freely accessible online portal for Swiss journals. Optimised for human usage, pages or entire articles are presented and issued in the form of PDFs on this portal. The OCR-recognised text is embedded in these PDFs.

Based on a sample of two architecture journals totalling around 380,000 pages the goal of this cooperative project is to verify the extent to which computer-linguistic methods can be used to both optimise OCR results and enhance the content of the text corpus. In doing so, the library is in charge of preparing and supplying the OCR text files, while the Institute of Computational Linguistics focuses on the subsequent data processing. On the one hand, the sample serves to improve the existing correction of OCR errors. On the other hand, the method used to recognise and link *named entities* – i.e. people and places – can be refined. The result should yield a machine-readable text corpus of the two journals where the OCR errors have largely been rectified and, where possible, the *named entities* have been identified

and linked. The advantages of this kind of text enhancement are obvious: not only do they facilitate searches within the text corpus; in particular, they also enable linkages with other data pools or retrieval systems.

The cooperative project *Thomas Mann Personal Library* between the Chair for Literary and Cultural Studies at ETH Zurich and ETH Library's Thomas Mann Archives is heading in a different direction. Thomas Mann's erstwhile personal library is one of the archives' most important holdings, containing numerous books with notes and markings made by the author's own hand. These marginalia are crucial to understand Thomas Mann's work. Funded by the Swiss National Science Foundation, the project's objective is therefore to digitise and transcribe all the handwritten traces left in his personal library and provide a state-of-the-art way for researchers to search for them.

The pure digitisation of the roughly 1,500 annotated books and brochures is a painstaking but established process. The academic requirements for recording and presenting the annotations and marginalia are what pose the actual challenges for the library. A core synergy arises with the platform e-periodica.ch. Individual system components of this platform are being expanded in such a way as to meet the specific project requirements. An editor facilitates the exact description of the annotations and side notes, and a custom web application will ensure the online presentation of the digitised texts with all their markings, handwritten annotations and the transcriptions produced until the completion of the project in the spring of 2019. Interesting possibilities for generating direct added value from digitisation via research partnerships also arise outside the *digital humanities*. Seismologists and astronomers, for instance, are interested in comprehensive, older, analogue series of measurements kept in the ETH Zurich University Archives, which belong to ETH Library. The Swiss Seismological Service located at ETH Zurich is concentrating on a series of seismological records from 1927 to 1955. Within the scope of the project, medium-strength earthquakes from this period are re-recorded using modern methods based on digitised seismograms. Once transformed, the historical series of measurements is channelled into the Seismological Service's modern data pool. A similar project co-initiated by the Solar Influences Data Center Analysis in Brussels involves re-recording solar activity according to modern standards based on over 22,000 digitised sunspot and protuberance records predominantly compiled at ETH Zurich between 1884 and 1949.

User participation

Digitised content not only creates added value and new possibilities in the field of research collaborations; the same also goes for modern forms of direct user participation. The most prominent example of this is crowdsourcing. This form of user participation, as e.g. Ridge demonstrated, is used highly successfully by a wide variety of institutions all over the world to enhance digital collections and raise their profile. [5]

At ETH Library, crowdsourcing has especially established itself as a key instrument in the Image Archive, which first used the model between 2009 and 2013.[6] In the first crowdsourcing project, former members of Swissair's staff volunteered to help index a selection of around 40,000 images from Swissair's photographic holdings. A small group of retired pilots and members of the ground and cabin crews were given their own login for the image database E-Pics BildarchivOnline (ba.e-pics.ethz.ch) and supplemented any missing or incomplete information on the aircraft types, places or people depicted.

In order to exploit the full range of specialist knowledge available via crowdsourcing, a freely accessible, easy-to-use feedback function was set up on the image database in late 2015 and advertised in various places under the banner of *Do you know more?* An article on the image database and its crowdsourcing function published in the *Neue Zürcher Zeitung* in January triggered an unimaginable response from the public and in the media.[7] In the space of a few days, hundreds of pointers on previously unidentified images came flooding in. Additional reports in the media, on the radio and on television increased the interest even further. By the end of September 2016 alone, the descriptions of almost 5,000 images could be improved thanks to nearly 7,000 pointers received – priceless added value that the Image Archive's seven-strong team could never have accomplished by themselves. The Image Archive's Community Management, which meanwhile is up and running, ensures that the crowd of volunteers remains dedicated for the long haul. Needless to say, this also includes thanking them for the pointers received and processing them swiftly. However, the weblog *Crowdsourcing at ETH Library: News and Experiences from the Community* (blogs.ethz.ch/crowdsourcing) launched in May 2016 is also important for communicating with the crowd. In about two posts per week, images that could not be classified geographically and/or dated are presented as an appeal, and images that have already been

---

[5] Ridge (2014)

[6] Graf (2016)

[7] Kälin (2016)

identified successfully thanks to the crowd are presented in more detail. Moreover, monthly statistics updates and lists of the top-ten volunteers are published on the weblog, and a small series of videos on the Image Archive's crowdsourcing and its most prolific volunteers is also distributed via this channel.

In the Image Archive's experience, active community management is important for the long-term success of a crowdsourcing campaign. Naming the person who provides the information and offering free access to the digitised images wherever possible have emerged as additional success factors. For most volunteers, seeing their name in the comments field on the Image Database containing the additional information from the crowd is an incentive. The fact that the majority of the roughly 400,000 currently digitised images are under a public domain mark or a CC BY-SA licence and can be downloaded freely and in the highest resolution caters for a mutual give and take. With its ongoing digitisation endeavours, the Image Archive is making an ever-increasing number of image sources freely accessible online. The growing amount of metadata improved through crowdsourcing also benefits all users.

The web-based integration of expert knowledge in the metadata for digital copies is not just restricted to the Image Archive at ETH Library, however. Over 1,100 historical maps were georeferenced in the space of a few weeks during a crowdsourcing project launched by the Map Collection, for example.[8] These digitised map sheets, which were already available on the platform e-rara.ch, were also rendered available for this purpose via www.oldmapsonline.org, the major international portal for historic maps, and linked to the tool Georeferencer (www.georeferencer.com), which enables volunteers to align historic maps with their modern counterparts by placing reference points. Thanks to this georeferencing, the digitised historic maps can be incorporated into modern geoinformation systems and re-used academically.

An entirely different input is expected from volunteers in the current project *e-manuscripta Volltext*, the aim of which is to expand the cooperative presentation platform e-manuscripta.ch to include a transcription module for handwritten sources. The project is headed by Zentralbibliothek Zurich and ETH Library is involved with regard to content and as a platform operator. The additional module should be available from the beginning of 2018 and facilitate simple transcriptions of letters, notebooks and other types of sources for registered users. In future, not only will the indexing information be used exclusively for searches on e-manuscripta.ch, but also a growing number of full texts produced by users.

---

[8] ETH Library (2017b)

Outreach

The example of georeferenced maps has already pointed to the further potential that digitised content harbours: the increased use of digital copies thanks to dissemination via various platforms. Thanks to the additional incorporation of historical maps in the major specialist portal oldmapsonline.org, the map sheets are where the specialist international audience is most likely to look for them.

The strategy of supplying digitised content via various channels and thus offering it where most users are plays a considerable role in drawing the attention of new customer groups to the institution and its holdings. This is particularly evident in the impressive example of the mass uploading of sub-collections from the image database E-Pics BildarchivOnline onto Wikimedia Commons. ETH Library decided on this parallel publication of image holdings at the end of 2015. The organisational and technical processes for batch uploads were subsequently implemented in a cooperative project with Wikimedia CH. The first productive uploading of 350 photographs with metadata onto Wikimedia Commons took place in mid-June 2016. By June 2017, around 37,500 digital copies had been uploaded.[9] An ever-increasing number of these images from Wikimedia Commons are being integrated in Wikipedia articles and reaching a global audience of millions via this route. Overall, more than 5.7 million page views were registered for all Wikipedia articles containing integrated images from ETH Library between May 2016 and May 2017 (spread among various language versions of the online encyclopaedia). For instance, photographs of the numerous foreign expeditions conducted by Swiss aviation pioneer Walter Mittelholzer (1894–1937) to the Middle East and Africa are found on pages of the English, German, French, Spanish and Arabic Wikipedia.[10]

It is not just the propagation of digitised content via several channels, but also novel ways of presenting digitised content that are ideal for addressing broader or even new user groups. This is where Explora (www.explora.ethz.ch/en), ETH Library's new storytelling platform, comes in. It presents topics in the form of stories in a compelling and visually attractive manner. If historical holdings play a central role, such as in the story *Mallow, iris, orchid. Scientific plant images over the centuries*, naturally a rich selection of existing digital copies

---

[9] Direct access to all the images uploaded to Wikimedia Commons by ETH Library via commons.wikimedia.org/wiki/Category:Media_contributed_by_the_ETH-Bibliothek.

[10] For the latest overview and statistics on Wikipedia articles with integrated Wikimedia Commons media files that have been uploaded by ETH Library, see tools.wmflabs.org/glamtools/baglama2/index.html.

is integrated in Explora.[11] In addition, however, supplementary multimedia content is also produced for the stories. Explora uses animations and video interviews with experts to bridge the gap to the present day and current research issues. Thanks to this strong visual focus, Explora also creates attractive added value for users of mobile devices and, not least, invites users to browse ETH Library's various presentation platforms.

Challenges and outlook

Digital humanities projects, crowdsourcing, storytelling: the possibilities, the potential in using digital copies, digitisation processes and online presentation platforms that have been created and established in recent years to the advantage of users and institutions, seem virtually inexhaustible. However, this wealth of possibilities is not without its fair share of challenges. From the institution's point of view, this does not just include defining the strategic spheres of activity, prioritising new projects and services, and providing the corresponding resources; the willingness to acquire new knowhow and respond flexibly to dynamic developments is also crucial.

With an eye toward the three areas outlined above, research partnerships probably demand the highest standards in the development of competence and knowhow. If researchers approach the library, their specialist concerns and research interests do not just have to be understood; it is far more a question of translating this input into specific process and application requirements. If the library seeks research partners for individual projects of its own accord, corresponding descriptions of the respective terminology need to be adapted. Key input for the preparation of the project with computer linguists, for instance, is owed to a member of ETH Library's staff who completed her degree in this field. In the start-up phase for Explora, on the other hand, storytelling workshops were conducted with external experts based on story canvases and other tools. The results of these workshops will form the basis for the next contributions to appear on this platform.

Responding swiftly to dynamic developments was – and still is – a central challenge for the Image Archive's crowdsourcing activities. Due to the surprisingly large response, answering and processing the wealth of pointers received as quickly as possible has become a key aspect of the day-to-day business for the Image Archive. Due in no small part to the Image Archive's consistent open-data strategy, corresponding resources could be provided: the possibility of downloading the majority of the digital copies freely and in high resolution led to a major

[11] Gasser (2017)

decline in user queries that needed to be processed manually. Instead of complex image orders, valuable indexing information from the population could now be processed.

What are the next steps? Further expansion is scheduled in all three areas. Combined with the research collaborations, the aim is to develop services and tools that expand ETH Library's portfolio for the long term via current and additional projects – whether this be in the form of enhanced text corpora, new web applications or selected additional measurement data from longstanding series of analogue measurements. As far as user participation is concerned, the Image Archive's crowdsourcing is to be expanded further. The next objective is to georeference oblique aerial photographs of Swiss landscapes. By the beginning of 2018, around 180,000 digitised aerial and landscape photographs will have been incorporated into the collaborative crowdsourcing platform sMapshot (smapshot.heig-vd.ch), where volunteers can position the images precisely on a virtual globe. This enables geocoordinates of the area depicted to be calculated. Not only is the project expected to generate valuable metadata; it should also boost the publicity and use of these image holdings even further.

Additional outreach is also facilitated in general by the increasing standardisation of the use of digital copies and their metadata across platforms. In the space of a few years, the International Image Interoperability Framework (IIIF) has significantly grown in importance. Accordingly, ETH Library is also working on enabling content-sharing for its presentation platforms via IIIF – as an additional but by no means last step towards optimising the use of its digital copies. The full potential certainly has not been exhausted.

References

ETH-Bibliothek (2016): Jahresbericht 2015 der ETH-Bibliothek. Zürich: ETH-Bibliothek. doi: 10.3929/ethz-a-004157606.

ETH-Bibliothek (2017a): EIDOS project description. Available at www.library.ethz.ch/en/About-us/Projects/EIDOS.

ETH-Bibliothek (2017b): Maps crowdsourcing project description. Available at www.library.ethz.ch/en/Ressourcen/Geodaten-Karten/Crowdsourcing-Projekt.

ETH Zurich (2014): Sammlungen und Archive der ETH Zürich: Strategie 2015 bis 2020. Available at www.ethz.ch/content/dam/ethz/main/campus/bibliotheken/Sammlungen-Archive_Strategie_2015-2020.pdf.

Gasser, Michael (2017): Mallow, iris, orchid. Scientific plant images over the centuries In: *Explora. A world of experience by ETH Library*. doi: 10.22010/ethz-exp-0001-en.

Graf, Nicole (2016): Sie wussten mehr! Vielen Dank! 'Offenes' Crowdsourcing im Bildarchiv der ETH-Bibliothek. In: Bienert A (Ed.) *EVA Berlin 2016*. Berlin: Staatliche Museen zu Berlin - Preussischer Kulturbesitz, 163–168.

Kälin, Adi (2016): Wer kennt die Berge, Orte und Fabriken? In: *Neue Zürcher Zeitung*, 18 January 2016. Available at www.nzz.ch/zuerich/wer-kennt-die-berge-orte-und-fabriken-1.18678913.

Mumenthaler, Rudolf; Voegeli, Yvonne (2005): Ohne Bibliothek keine Wissenschaft. In: ETH-Bibliothek Zürich (Ed.) *Blättern & Browsen – 150 Jahre ETH-Bibliothek.* Zürich: ETH-Bibliothek.

Ridge, Mia (2014): Crowdsourcing Our Cultural Heritage. Farnham: Ashgate.