Data Life Cycle Lab
Key Technologies

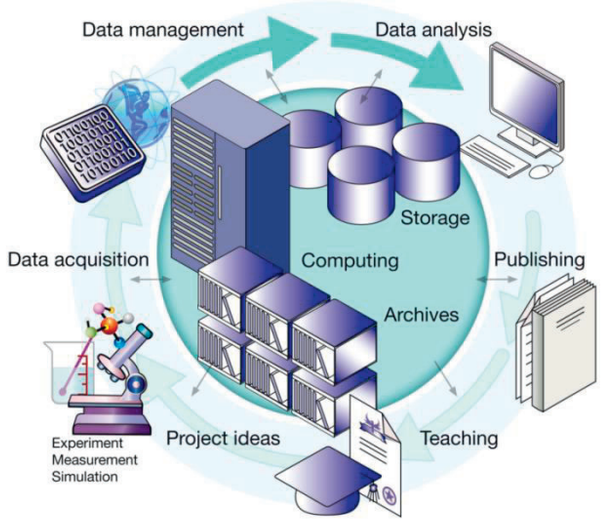
Status 2013

Big Data in Science

LSDMA

EDITORIAL

Data exploration is one of the key challenges in many scientific projects. Huge volumes of measurement and analysis data are produced in very short periods of time. Common technologies to manage, analyze, visualize, archive, and share the data in its whole data life cycle (see figure) hardly exist.



The introduction of Data Life Cycle Labs (DLCL) in the portfolio extension “Large Scale Data Management and Analysis” (LSDMA) [LSDMA] of the Helmholtz Association of research centers in Germany enables new capabilities to develop jointly cross-disciplinary data technologies. Within the Data Life Cycle Labs domain and data scientists work closely together to optimize various stages of domain specific data life cycles. Additionally, a Data Services Integration Team extends the Data Life Cycle Labs by developing generic federated data services.

In this brochure the Data Life Cycle Lab “Key Technologies” and related work in Big Data in Science are presented by the partners Karlsruhe Institute of Technology, the universities of Heidelberg, Mannheim, and Dresden. It includes a variety of data intensive scientific projects, their special demands with respect to large scale data management and analysis as well as envisaged data technologies:

- *Instrumentation of and research on the world’s fastest tomography beamline,*
- *lightoptical microscopy of tiniest structures in nanometer scale,*
- *cutting-edge high throughput microscopy,*
- *the beauties of medieval manuscripts and digital humanities,*
- *a novel ultrasound computer tomography system.*

The indicated authors are responsible for the content, respectively.

We want to express our thanks to all contributing scientists, the LSDMA management at Steinbuch Centre for Computing (SCC) at KIT, the German Helmholtz Association and the German Federal Ministry of Education and Research.

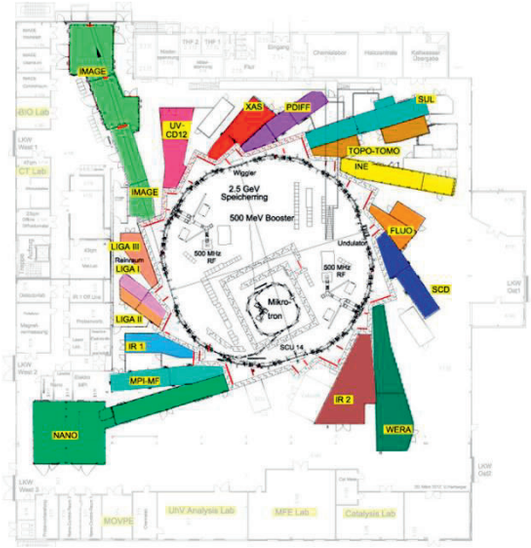
Karlsruhe Institute of Technology, September 2013

Rainer Stotzka

DATA MANAGEMENT AT ANKA

Halil Pasic, Rainer Stotzka, Thomas Jejkal, Xiaoli Yang — KIT, IPE
Wolfgang Mexner, David Haas — KIT, ANKA

ANKA is a synchrotron facility located at KIT consisting of an accelerator ring, and attached experimental stations (beamlines) utilizing the light produced by the accelerator. The experimental stations provide specialized setups and broad range of applications for national and international users.



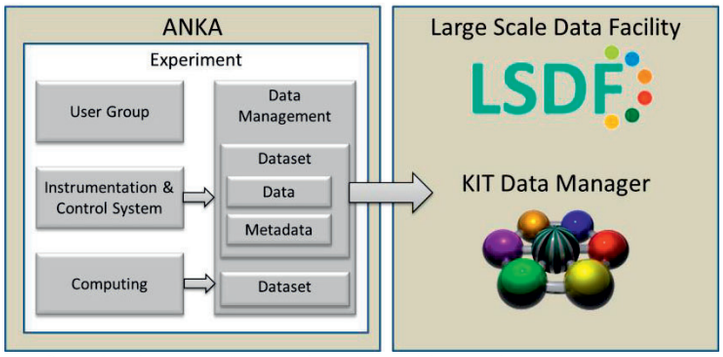
Schematic presentation of ANKA, experiment hutches in color.

One of the highlights is the Ultra Fast Tomography (see next pages) using the X-ray part of the light to provide high resolution four dimensional images (a series of attenuation volumes with subsequent volumes corresponding to subsequent points in time). Ultra Fast Tomography is used to get new insights in dynamic biological and physical processes.

Some of the most important functional requirements towards data management at ANKA are:

1. **Automated data management and federated data access:** Users do not need to be involved into technical details of the data management and data policies.
2. **Flexibility**: The continuous improvement of the experimental setups is an important part of ANKA research.
3. **Large scale data:** ANKA produces multiple petabytes of data yearly.
4. **Long living data and metadata:** Data which lead to publications needs to be archived for at least 10 years. Therefore it needs to be extended with metadata.
5. **Heterogeneous data:** The beamlines support many experiments and users.

A new concept to manage the data life cycle including transparent workspaces was required and is being implemented [ANKA1]. It is based on two connected data management infrastructure projects at KIT. The Large Scale Data Facility (LSDF) supports and enables scientific projects by providing reliable large scale storage and processing infrastructures. The KIT Data Manager (KIT DM) extends the LSDF by a data service architecture with additional data and metadata services allowing federated data ingest and access.



Data in the experiment domain and beyond.

In the ANKA experiments both primary and derived data are produced. The primary data is created during measurements by the instrumentation and the control system, parameterized by the user. Derived data is calculated from primary data using on-site computing infrastructures.

To deal with the heterogeneity of the data and to manage the data based on policies, data managed in the experiment domain is organized in datasets from the very beginning. A dataset is a basic unit of managed data and consists of one or more unique identifiers, data and metadata. Metadata is captured automatically from the instrumentation and control system, if possible, and needs to be extended by the user. A variety of tools supports the administration of the experiment data life cycle. In the tomography beamlines the standardized data format NeXuS as well as proprietary formats are used allowing the application of proprietary analysis tools.

The results of a successful experiment are archived in the LSDF using the KIT DM. External users can access and process their experiment data remotely.

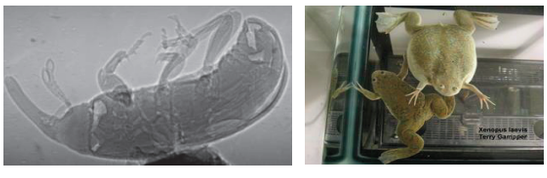
DATA ANALYSIS FOR ULTRA FAST X-RAY TOMOGRAPHY

Xiaoli Yang, Rainer Stotzka, Thomas Jejkal — KIT, IPE
Tomy dos Santos Rolo, Thomas van de Kamp, Julian Moosmann, Ralf Hofmann — KIT, IPS

The Ultra Fast Tomography system at the ANKA Synchrotron Light Source at KIT allows to study moving biological objects with high temporal and spatial resolution. The resulting amounts of data are challenging in terms of reconstruction algorithm, automatic processing and computing:

- Fewer projections from Ultra Fast tomography lead to reconstructions with artefacts using standard algorithms
- Laborious manual process for data analysis
- Large amounts of data sets and metadata
- Computational expensive

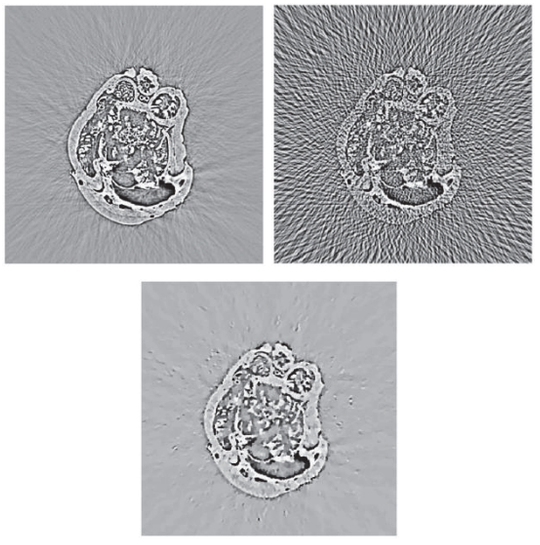
Ultra Fast Tomography beamline at ANKA.



Moving biological objects: a living bug (left) and African clawed frogs (right).

Sparse Reconstruction

Sparse reconstruction in computed tomography (CT) refers to an estimation of a numerically accurate tomographic image if the projection data is not sufficient for exact image reconstruction according to the Nyquist-Shannon sampling theorem. We developed a new algorithm ART-CS combining the algebraic reconstruction technique (ART) and the compressive sampling theory (CS) optimizing the images based on total variation. Compared to the standard algorithms the results show high quality images even if only 60 instead of 1500 projections are used.

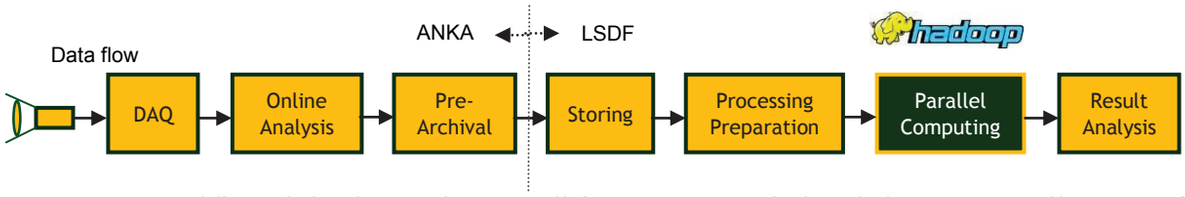


Top left: standard reconstruction with 1500 projections; Top right: standard reconstruction with 60 projections showing artefacts; Bottom: ART-CS reconstruction using 60 projections.

Parallel Computing

The reconstruction of a full volume on a standard workstation using ART-CS requires 11 hours of computing time. For near real time processing a HADOOP cluster connected to the Large Scale Data Facility (LSDF) is used. After data ingest a reconstruction workflow is started automatically processing the MATLAB programs in parallel. By this the computation time is reduced to 6 minutes [ANKA2].

ART-CS and the parallel computation using the LSDF infrastructure are general concepts and can be used for many other tomography systems.



An automatic workflow of the data analysis: parallel computing, marked in dark, proves its efficiency and effectiveness; speedup for a data set reconstruction goes up to 120.

ULTRA FAST X-RAY TOMOGRAPHY TO REVEAL THE ANATOMY OF SMALL ORGANISMS

Thomas van de Kamp, Tomy dos Santos Rolo — KIT, IPS
Andreas Kopmann — KIT, IPE

Scientific motivation

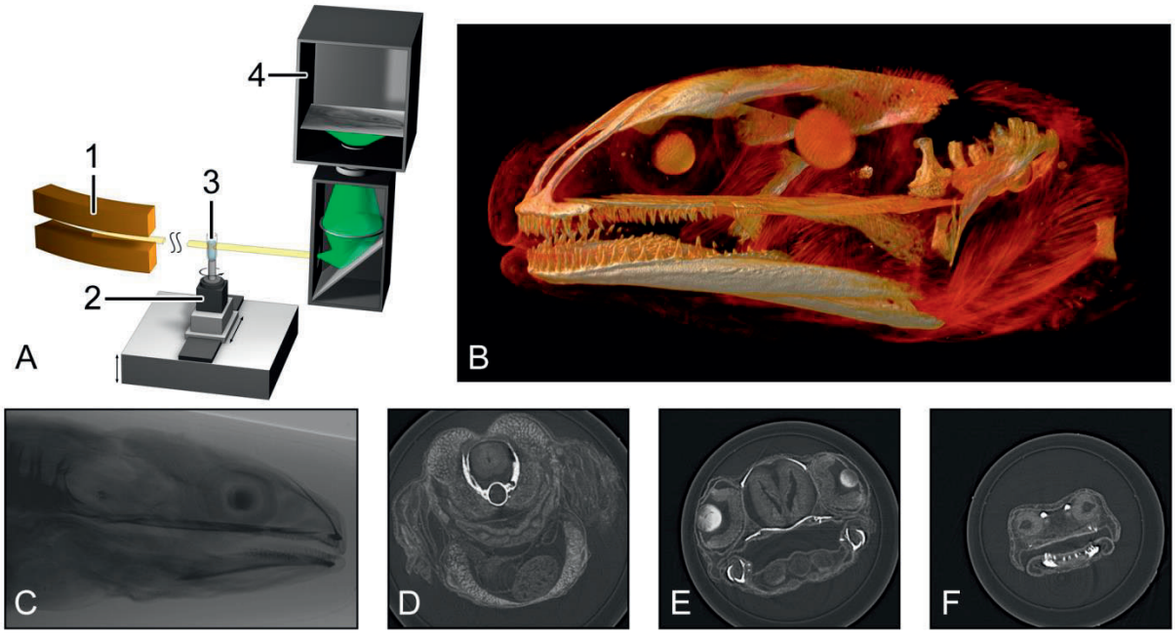
X-rays and tomography provide the opportunity to visualize internal structures of optically dense materials in 3D. The invention of synchrotron-X-ray-microtomography was the onset of a new era of morphological research on microscopically small animals. Analyzing such 3D-data is time consuming and technically challenging. Especially the automation of classification processes needs close cooperation of biologists and image processing experts.

Technologies

Right now, a new high-speed-tomography setup (UFO) is built at the ANKA synchrotron in Karlsruhe (KIT) [UFO1]. This setup will enable unrivaled opportunities of high-throughput measurements and 3D/4D-tomographic imaging of dynamic systems and living organisms. However, this setup results in such large amounts of data that technical limitations will be reached regarding data acquisition, storage, organization and analyses. Therefore, users will no longer be able to process the resulting data at their home institutions, even with highly equipped computers. Hence, a strong integration of knowledge from biology, image processing and data management is needed to enable this new and fascinating high-speed-option for regular users in the future. We aim to establish and standardize measuring parameters for the UFO setup to meet the needs of a broad range of biological research. This will be achieved by optimizing data acquisition and processing, semi-automation of data analysis (reconstruction and segmentation/classification) and the creation of an online-portal providing easy access, stereoscopic visualization and semi-automatic analyses of the data using cloud technologies.

Data

Data sets typical consist of 20-50 3D tomograms with 10-50 GB each. Metadata will be retrieved from the ANKA user management system and the beamline control system.



Fast synchrotron X-ray microtomography. A: Experimental setup for ultrafast X-ray microtomography showing bending magnet (1), rotation stage (2), specimen (3) and detector system (4). X-rays travelling from left to right pass through the sample mounted on a rotary stage with two translational degrees of freedom. A detector converts X-rays to visible light that is subsequently recorded by a camera. B: volume rendering of a newt larva. C: Radiograph of the same sample. D - F: Slices of the reconstructed volume.

FAST ANALYSIS OF IMAGE STACKS IN OPTICAL NANOSCOPY

Nick Kepper, Michael Hausmann — Uni-HD, KIP
Jürgen Hesser — Uni-HD, ERO

Scientific Motivation

Light microscopy is a routine imaging technique in biological and medical research and diagnosis. Although nowadays instrumentation has made substantial progress concerning imaging quality and speed, there is still a gap in resolution between light microscopy (~200 nm) and electron microscopy (~10 nm). This so far missing scale range would however open new insights into the nanocosm of a cell and its sub-cellular structures [NANO1].

Localization nanoscopy, being a candidate to fill this gap, is a novel technique overcoming resolution limits due to diffraction. During the last decade several setups have been developed and used to answer interesting and challenging questions in the field of cellular biology and molecular biomedicine [NANO2].

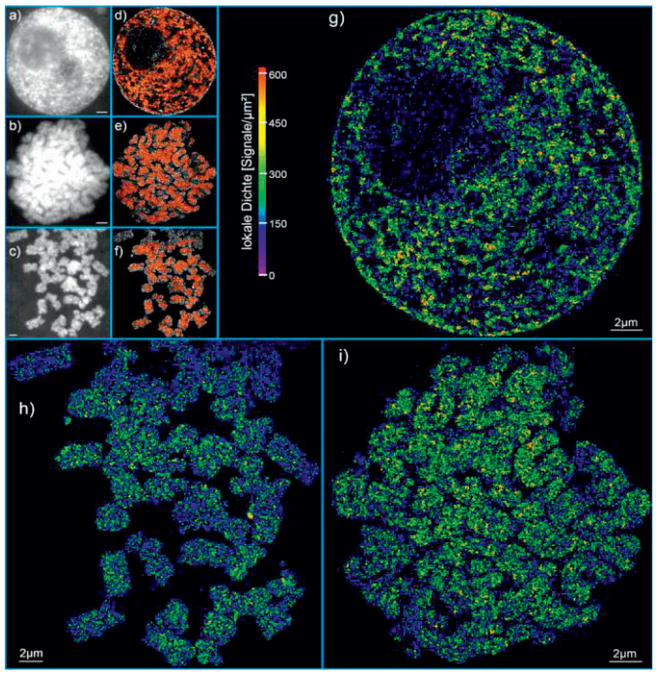


Figure 1: Histone H2B distribution in HeLa cell nuclei and (pro-)metaphase chromosomes [NANO1].

Instrumentation

For localization nanoscopy standard microscopic optics and fast imaging systems are required. The principle of the so far developed techniques depends on optical isolation and separation of individual dye molecules by their spectral signature. The embodiment (SPDM = Spectral Position Determination Microscopy) used in our collaboration makes use of dye molecules for specific labeling of cellular sub-structures that are able to undergo so called reversible photo-bleaching which results in stochastic molecular blinking. Taking a huge time stack of images (~1000 frames) the switch off/on of each molecule can be detected and the molecular coordinates can be determined precisely (in the range of nm). Hence, distances between dye molecules can be calculated in the ten nm range and thus sub-cellular structures can be visualized and measured also in 3D conserved cells or even under vital conditions.

Examples

In the following two typical examples will be explained: In Figure 1 an example of a cell nucleus and (pro-)metaphase chromosomes are shown. a) – c) show the wide field microscopic images; d) – f) present the merged images from the time stack of SPDM displaying thousands of individual molecules by a color dot. In g) – i) these images are enlarged and coded according to the numbers of next neighbors so that structural information can be elucidated [NANO3]. Such chromatin 2D/3D nanostructures are of importance to understand chromatin rearrangements during repair processes of DNA after exposure to ionizing radiation. This information is used to create and validate a consistent architectural model in the field of radiobiology.

The images of Figure 2 show on the left an overlay of a standard wide-field image (green) and the result of localization imaging (red) of a membrane section of a breast cancer cell where the Human Epidermal growth factor Receptor 2 (Her2/neu, a typical breast cancer marker) is specifically labeled. The right image shows the result of localization imaging which is obtained from a time series of 1000 image frames (979 x 816 pixels, 150 ms per image). Here, each point represents a single fluorochrome respectively antibody attached to a receptor molecule. In contrast to the wide-field image that does not allow to identify any detailed nano-structural information about the spatial arrangement of the antibodies/receptors, the resulting localization image reveals details of the formation of receptor clusters or linear arrangements of receptors (inserts) which can be correlated to dimerization induced functional activity [NANO4].

These examples indicate the huge progress going along with localization nanoscopy. However the volume of the data is drastically increasing by orders of magnitudes requiring novel approaches of managing, archiving and analyzing.

Technologies

From the examples shown above we assume the digital volume of one cell nucleus of about 20 µm diameter with a resolution of approximately 10 nm is about 32 GB per channel of color. In larger screening experiments the limit of one PB data volume is thus reached easily. For the highly sensitive analyses and structure elucidation, very complex and highly variable algorithms have to be used to avoid artifacts and to find out structural re-arrangements. This includes iterative variation based denoising and deblurring techniques. Still the data is saved and worked on in an ad hoc manner, which with serial computation systems leads to extremely long processing times and a limitation of the selectable volume size due to limitations in the computer memory. The data rate created by a nanoscope is in the range of up to GB/s depending on the size of the detected region of interest and the dimensionality (2D/3D) required for scientific investigations.

Actual algorithms and techniques have been developed for a PC basis without usage of techniques for parallelization. This strongly limits the handling of large data sets as being necessary in biological research and medical diagnosis especially if a serious significance of statistics is required (i.e. if a large series of cells have to be evaluated). Here, we develop a pipeline for parallel data analysis. Variation based methods need, with parallel analysis of the data, a synchronous update of all analyzed regions, which will be realized with message passing. The access to the data has to be self-explaining for the user and has to fulfill the rules for storage of the DFG for several years.

Data

Due to the increasing data volume and the complexity of the analyses of nanoscopic image data, new and more advanced techniques for image analysis and storage are needed. The goal for the image acquisition is an isotropic resolution of 10 nm of point patterns, which, in case of a nuclear or cellular diameter of 20 µm results in the magnitude of 32 GB per color channel. The major challenge is to reconstruct the point pattern free of any bias, which is very demanding of computational power with this amount of data. Moreover from such point patterns continuous structures have to be calculated and interpreted. Screening experiments can easily need data volumes greater than one PB.

The data can be analyzed in parallel, since the information in pictures after the calculation of the positions of the fluorescence molecules, is provided in a grid structure. It is planned to use variation based techniques, which put into effect a-priori-reconstruction, as on the one hand it follows the physical image, and on the other hand also the sparseness as a priori knowledge about the structures is used (images of these structures will reach high compression rates). In this case, an algorithm in parallel will be developed, which yields good performance on computer clusters. The synchronization between the nodes will be realized by message-passing.

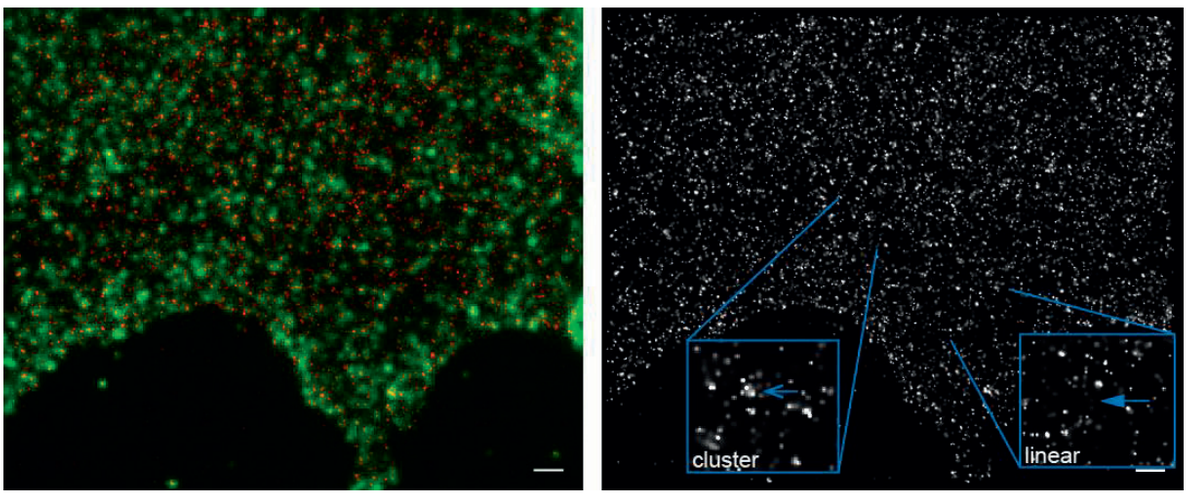
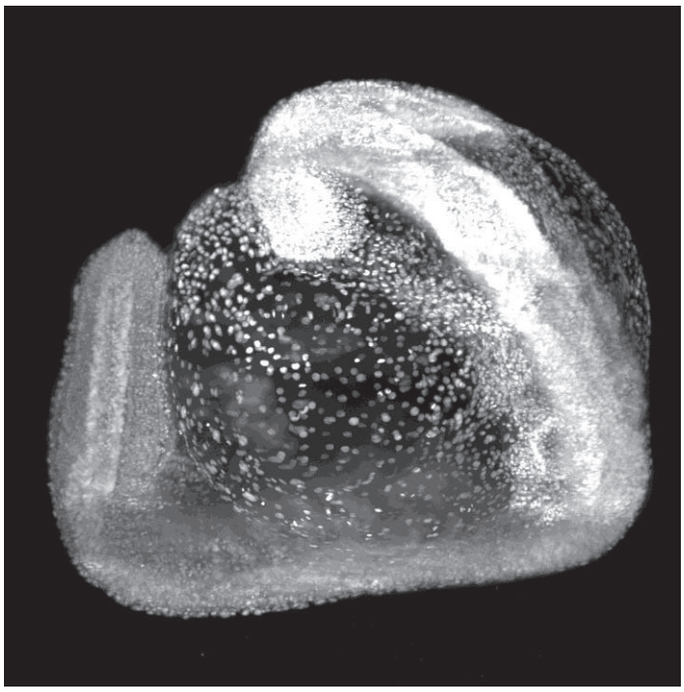


Figure 2: Image section of the membrane of a breast cancer cell after specific labeling of the Her2-receptors by means of fluorescence labeled antibodies. (courtesy J. Neumann, Kirchhoff-Institute for Physics, University of Heidelberg).

LIGHT SHEET MICROSCOPE (LSM)

Andrey Kobitskiy, G. Ulrich Nienhaus — KIT, APH
Jens C. Otte, Masanari Takamiya, Uwe Strähle — KIT, ITG
Johannes Stegmaier, Ralf Mikut — KIT, IAI



Maximum intensity projection of an image stack depicting a zebrafish embryo at 24 hours post fertilization.

Scientific Motivation

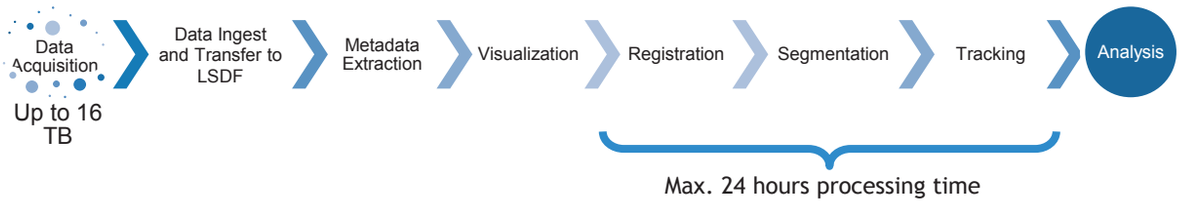
Analysis of cell movement during vertebrate embryogenesis requires a microscopy apparatus capable of 3D image recording with high spatial and temporal resolution. To achieve an individual cell tracking over the entire embryogenesis it is necessary to keep the high acquisition data rate constant over prolonged time periods of 24 hours or more, see survey in [LSM1]. Moreover, for the correct data interpretation it is important to have good statistics on many biological samples.

Therefore, a novel optical system based on digital scanned laser light-sheet microscope has been built at the APH KIT Campus South. The microscope is capable of image recording with the speed exceeding 10^8 voxels/s. Thus, the full 3D stack, covering the field of view of 1039×876 μm^2 laterally and 1000 µm axially with the resolution of 2560 and 2160 pixels and 500 frames, respectively, is acquired within 25 seconds. This image acquisition rate was kept stable for more than 24 hours, while the total acquisition time of a dataset was limited only by the available local disk storage space of 16 TB.

Recently, the microscope has been utilized to record more than 24 datasets within six weeks of the first 12-16 hours of zebrafish embryo development with the total data size above 250 TB. To achieve high sample throughput recorded data were transferred after every measurement to the LSDF (KIT Campus North) over the 10 GE network with the rate exceeding 400 MB/s. To manage such high data rate a special protocol, gridFTP, optimized for the fast transfer of big files has been used (see also next page).

Data Analysis

Due to tremendous amount of microscopy data time-efficient algorithms for the automated cell nuclei identification has been developed at the IAI KIT Campus North. Thus, by utilizing high parallelization during segmentation a complete data set of 10 TB from a single experiment can be processed in less than 24 hours on the Apache Hadoop computer cluster (KIT Campus North).



Data analysis workflow.

DATA MANAGEMENT TOOLS FOR LSM

Volker Hartmann, Francesca Rindone, Thomas Jejkal, Lukas Niedermaier, Rainer Stotzka KIT, IPE

Data Ingest Client

As the LSM data is too large to be stored locally at the experiment, it has to be ingested efficiently into the Large Scale Data Facility (LSDF). The Data Ingest Client automates the ingest process by collecting the data, providing additional (administrative and content) metadata and transferring the data using the Abstract Data Access Layer API (ADALAPI).

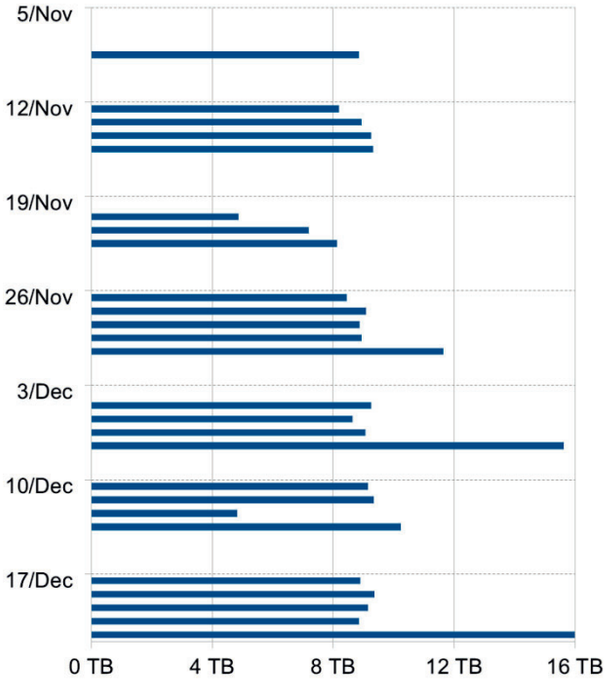
ADALAPI

The ADALAPI is a library implemented by the KIT/IPE to transfer data supporting multiple protocols.

Features:

- Authentication
- Multi-Protocol Support
  - GridFTP
  - GSI-SCP
  - WebDAV
  - HDFS
- Rebuild Canceled Transfers

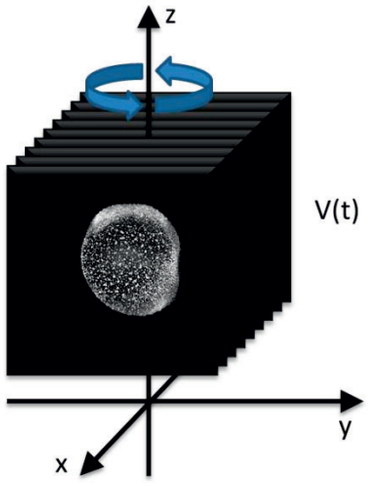
In case of LSM GridFTP is used due to the big data files and the required high transfer rate. An average transfer rate of more than 400 MB/s is reached for all data. In total more than 250 TB of data have been transferred to the LSDF in six weeks.



Data transfer per day over a period of six weeks.

Volume Visualization

Directly after the data ingest a workflow is started automatically to compute a 3D visualization of the temporal development of a zebrafish embryo. For the visualization each image stack V(t) of the 16 TB dataset has to be accessed, scaled, rotated, and a maximum intensity projection is computed. On a single workstation this process requires approximately 100 hours of computing time. A parallel computing infrastructure based on HADOOP and attached directly to the LSDF delivers the computing and data throughput capabilities. The result is a movie with less than 50 MB allowing a quick visual quality check of the experiment.



Rotation and projection of a time-dependent image stack V(t).

Data Access and Generic Search

Identification of a single dataset hidden in large storage and archiving systems requires performant search tools using metadata. All data stored in the LSDF is linked to metadata. Three different types of metadata exist:

- Administrative metadata
- Content metadata
- Structural metadata

As the administrative metadata has a similar structure for all data stored in the LSDF, a simple search for this type has been implemented as a first step. Content specific metadata is extracted automatically from the data, stored and propagated via OAI-PMH to a search engine.

ADVANCING CUTTING-EDGE BIOLOGICAL RESEARCH WITH A HIGH-THROUGHPUT WORKFLOW

Richard Grunzke, Ralph Müller-Pfefferkorn TUD, ZIH

Automatic Microscopy in Biology

The Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) in Dresden conducts research on the molecular mechanisms of absorption and transportation of molecules in cells. To examine the underlying cellular processes parts of cells are highlighted. For such an experiment genomes are prepared on plates (usually many) in small wells (up to 384 per plate). Each of the wells contains a slightly different sub-experiment. The preparation of the plates can take up to several months. The reactions in the cells create light emissions which are observed with an automatic microscope. Several images (with a size of up to 10 MB) are taken at different depths and at ten different positions in each well (Figure 1).

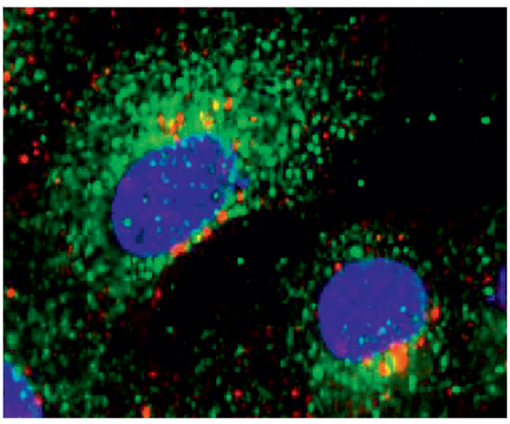


Figure 1: Picture taken by automatic microscope showing highlighted parts of cells.

By now several tens of millions of images were acquired which total amount of more than 100 TB. The images are analysed by the image analysis suite Motion Tracking, developed at MPI-CBG. High ranking publications, for example in the Nature magazine [BIO1], are the result. In the future, three-dimensional and time-resolved movies are planned, which will cause a vast increase in storage and analysis demands.

Image Analysis Workflow

The workflow is as follows [BIO2]. It is seamlessly integrated into the working environment of the biologists (Windows-based image analysis suite).

1. Images are acquired by the automatic microscope in large numbers.
2. The images are stored on a file server in distinct data sets.
3. The biologist selects files and loads them with the image analysis software. The templates for the job description files and run scripts are created by Motion Tracking including the paths of the input files.
4. The image for each job is synchronized to an iRODS server. Only new files are transferred. The files are stored in the same data set directory structure in iRODS as the original data set on the file server.
5. A UNICORE job is submitted via the UNICORE command line client UCC.
6. A UNICORE/X manages and forwards the job to the cluster.
7. The job is scheduled on the destination cluster and the run script is executed.
8. The script pulls the input files from the iRODS server. It authenticates via a proxy certificate created by UNICORE in the job directory on the cluster.
9. The run script executes Motion Tracking on the HPC system to analyse the image.
10. After the analysis is finished the results are put back into the iRODS server.
11. Motion Tracking regularly checks for new results and fetches them via the iRODS irsync.
12. Motion Tracking checks the results for correctness.
13. Motion Tracking saves the results by integrating them into the original data set.

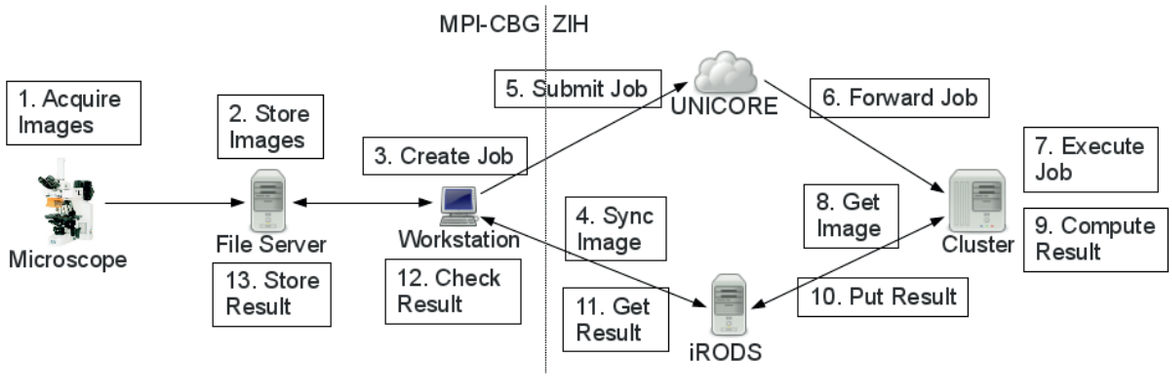


Figure 2: Image analysis workflow.

dawa - DATA WEB APPLICATION

Francesca Rindone, Danah Tonne, Rainer Stotzka KIT, IPE

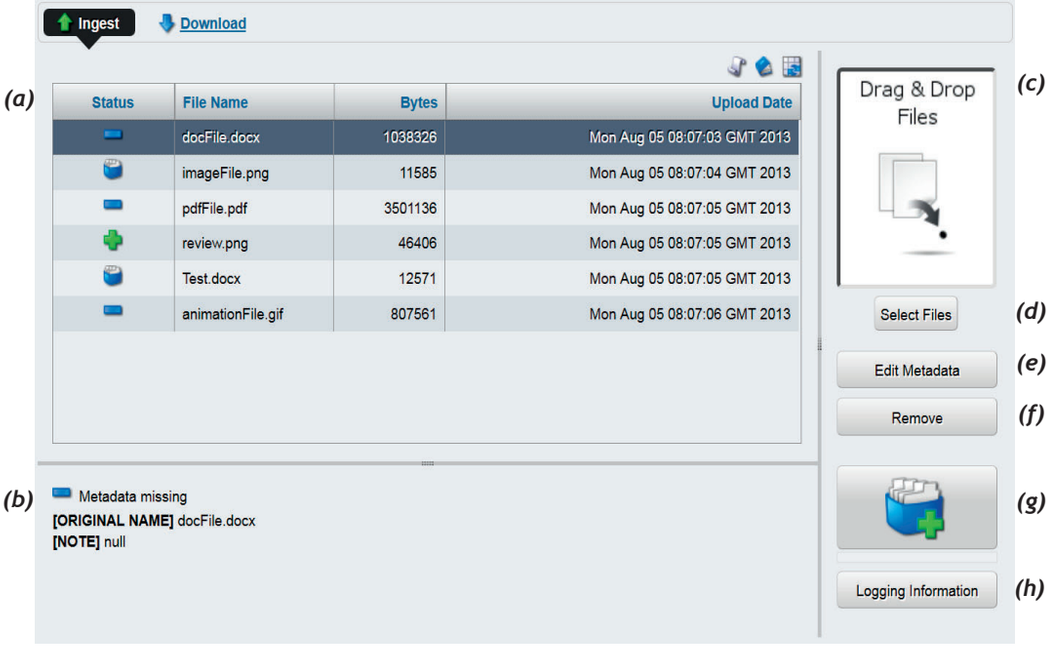
In the last few decades the production of data increased in nearly every research institution due to growing technical capabilities. In view of ever-increasing data amounts, in most different communities the question where to store big data in an easy, but secure way arose. For an easy-to-handle and secure long-term storage, a user-friendly web interface, known as **Data Web Application** dawa, has been implemented for the DARIAH project.

The Data Web Application is a Vaadin-based web application which can be integrated as a portlet in various Liferay portals.
In the context of dawa a **digital object** describes an object composed by data and descriptive metadata. Compared to the data, the descriptive metadata is not a mandatory component of a digital object. Thus, the simplest form of a digital object is a data file. For a systematic and detailed search for digital objects via higher level services, the adding of metadata is recommended.

The two main functionalities are the ingest and the download of digital objects to and from a storage system. Basis for using dawa as a client service in front of a storage system is a RESTful storage API which allows the communication between dawa and the storage system. A running dawa version, using the DARIAH Storage API [DARIAH1] as storage service, is currently available in the DARIAH Developer Portal.
The **ingest** process, executable on dawa's tab sheet *"Ingest"* (see figure), is defined by three steps:

1. Creation of an digital object by uploading data (and corresponding descriptive metadata).
2. Ingest to the storage system via the integrated storage service.
3. Assignment of a persistent identifier (PID, Handle service, GWDG, Göttingen, Germany) to the ingested digital.

The **download** of digital objects can be executed on dawa's tab sheet *"Download"*. Via the PID all information about a digital object can be requested.



Ingest View:

- *(a) Table - List of all digital objects planned to be ingested.*
- *(b) Output Panel - Displays all information for the use.*
- *(c) Upload Field - Uploads data files which have been dragged and dropped here.*
- *(d) Button - Uploads data files via a file dialog which allows the navigation through the file system.*
- *(e) Button - Opens the metadata editor for adding metadata describing the corresponding data file.*
- *(f) Button - Deletes table entries.*
- *(g) Button - Ingests digital objects to the storage service and creates a PID for each ingested object.*
- *(h) Button - Shows information logged during the ingest process.*

DARIAH – A EUROPEAN RESEARCH INFRASTRUCTURE

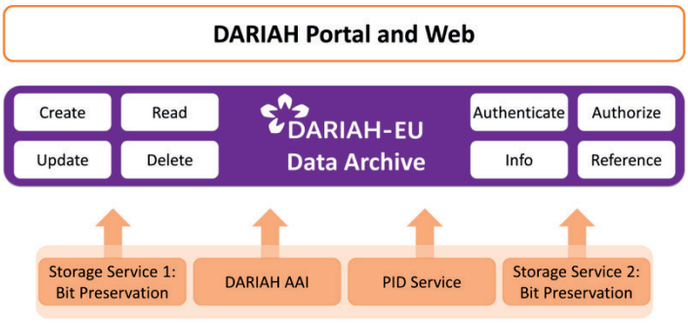
Danah Tonne, Rainer Stotzka, Francesca Rindone KIT, IPE

DARIAH-DE
Digital Research Infrastructure for the Arts and Humanities

Due to new digital methodologies and technologies the humanities research process has become more and more data-intensive in recent years. DARIAH (Digital Research Infrastructure for the Arts and Humanities) was therefore initiated to build up a sustainable research infrastructure for direct support of the scholars. DARIAH is one of the few humanities and socio-scientific projects placed on the ESFRI roadmap (European Strategy Forum on Research Infrastructures) and is based on national contributions from the participating countries. The German part DARIAH-DE is led by *Göttingen State and University Library* and unites 17 partners including various humanities disciplines, computer sciences and data centers.

Digitized codices are just an example of heterogeneous humanities research data (Public Library and Archive of Trier, Hs 1108/55 4° 6v and 7r).

DARIAH covers a wide range of topics, from the core infrastructure with generic or community-specific services to ontologies, meta data and license recommendations to training and education. An essential component of the infrastructure is a long-term storage serving the wide variety of disciplines in the digital humanities and accounting for their special requirements. The data from various projects and scholars differs in size (a few kilobytes for a text file containing a letter or several gigabytes for a film record of an opera), quantity (a few image files of a rare and valuable manuscript up to several millions of image files of a whole library) and type as there is a variety of different formats for text, image, audio, and movies. In addition to the data itself the scholars' diverse expectations need to be considered while designing and implementing services which support storing their data as convenient as possible.



Overview of the DARIAH Storage Architecture.

DARIAH provides a couple of stand-alone services usable for a sustainable storage of research data. The Storage Service [DARIAH1] offers a long-term and redundant file storage and ensures data integrity on bit-stream level. An identifier for this data needs to remain stable to be for instance cited in scientific publications, the Persistent Identifier Service therefore provides methods to create persistent identifiers (PIDs) and assign them to the data. As not all data is open to the public, for instance if personal data is contained, it has to be carefully protected from unauthorized access by the integration of an Authentication and Authorization Infrastructure (AAI). Dawa, the Data Web Application (see previous page) is the first service integrating these components.

Due to the heterogeneous research data and the long period of usage it is essential to create a modular infrastructure addressing various services on demand. Including standards and standardized interfaces simplifies future adoptions and exchange of underlying services. Nonetheless it is a prevailing field of research how to provide a sustainable and comprehensive research infrastructure for the arts and humanities.

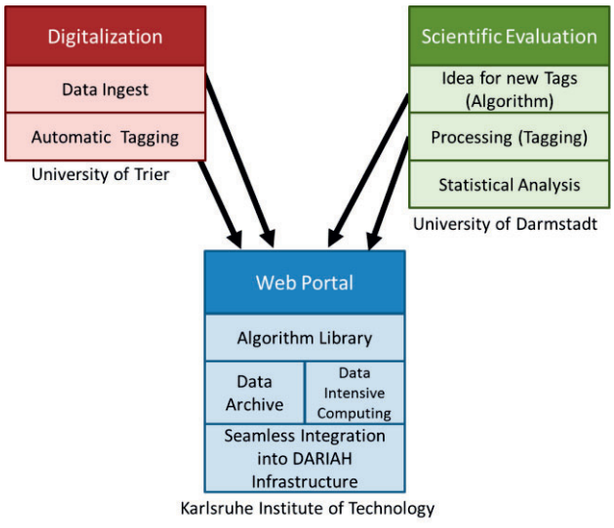
eCODICOLOGY - ALGORITHMS FOR AUTOMATIC TAGGING OF MEDIEVAL MANUSCRIPTS

Swati Chandna, Danah Tonne, Rainer Stotzka — KIT, IPE
Hannah Busch, Philipp Vanscheidt, Claudine Moulin — Uni-Trier
Celia Krause, Andrea Rapp — TU-DA

An essential component of historical research lies in the analysis of digitized medieval manuscripts which is reaching its technological limits due to lack of digital methods and algorithms for the analysis of these manuscripts. The BMBF-funded joint research project eCodicology uses the library stock of roughly 500 medieval manuscripts which have been written and collected in the library of the Benedictine Abbey of St. Matthias in Trier (Germany). The manuscripts were digitized and enriched with bibliographic metadata within the scope of the project "Virtual Scriptorium St. Matthias" (http://stmatthias.uni-trier.de/). Based on the provided images, the purpose of eCodicology is the development, testing and optimization of new algorithms for the identification of macro- and microstructural layout elements (see figure below) in order to further enrich their metadata.

Page size, image size, text size, initials.

The digitization of the manuscripts and the creation of a metadata schema and models for the XML files according to TEI P5 take place at Trier. After ingesting the digitized images into a data repository, they are processed at Karlsruhe [eCOD1]. Specific algorithms are adapted or designed and developed for the identification of macro- and microstructural layout elements like page size, writing space, number of lines, number of columns, proportion of text and pictorial space etc. The following scientific evaluation and a statistical analysis of the manuscript groups are performed at Darmstadt.



Collaboration within eCodicology.

A software framework tied to a web portal will automate the data analysis workflows. It is designed generically to process a great amount of image data with any desired algorithm for feature extraction based on the components ImageJ and MOA/Weka. Assuming a computing time of one minute per page, the one-time processing of the Virtual Scriptorium with a total of 170,000 pages would approximately take four months. Since algorithm development is a highly iterative task it is inevitable to utilize a cluster for data intensive computing. As a result, the hidden relationships of around 500 medieval manuscripts can be detected automatically and a database of objectified, reproducible and at micro level differentiated features will be created.

The software framework itself will be integrated as a service into the DARIAH infrastructure to make it adaptable for a wider range of documents and communities. Thus, eCodicology can show the potential of computer-aided methods by providing algorithms for the automatic and convenient tagging of medieval manuscripts.

3D ULTRASOUND COMPUTER TOMOGRAPHY

Nicole Ruiter, Michael Zapf, Torsten Hopp, Robin Dapp, Hartmut Gemmeke KIT, IPE

Motivation

Breast cancer is the most common cancer among women worldwide. In the project "3D Ultrasound Computer Tomography" (USCT) a new imaging method for early breast cancer detection is being developed. It promises three-dimensional images of the breast with high spatial resolution and specificity for breast cancer diagnosis. The aim is the detection of tumors with an average diameter of less than 5 mm to improve the survival probability of the patients.

Figure 1: Current 3D USCT prototype which is used for first clinical studies.

Technologies

The imaging principle is based on several thousand of ultrasound transducers, which surround the breast in a water bath acting as coupling medium. The patient lies in prone position on a patient bed, which also contains the signal acquisition hardware (Figure 1).

While one transducer emits ultrasound into the water bath respectively the breast, all other transducers receive the transmitted and reflected signals. By applying this acquisition step for all emitter-receiver combinations, a complete measurement for one breast creates up to 40 GB of raw data, which is currently acquired in approximately 4 minutes. For data acquisition the in-house developed multi-purpose hardware [USCT1] is used.

From the acquired signal data, high resolution volume images are reconstructed. USCT is able to provide qualitative as well as quantitative imaging of the breast. The applied reconstruction algorithms for reflection and transmission tomography are highly computational intensive due to the large amount of data. Therefore – besides the optimization of the algorithms for improved image quality – one of our key topics is the acceleration of the reconstruction methods using CPUs and GPUs.

After reconstructing the images (Figure 2) are prepared for diagnosis using clinical standards like DICOM. Furthermore a dedicated viewer software and an underlying data management has been developed. Especially for clinical studies, the raw data as well as the image data is subject to long-term archival according to the medical regulations.

Competences

For development of the USCT prototypes we apply the joined competences of our institute: Ranging from the mechanical design, development and production of ultrasound transducers, layout of analog and digital electronics to the development of the reconstruction algorithms. In addition we are cooperating with a number of national and international partners.

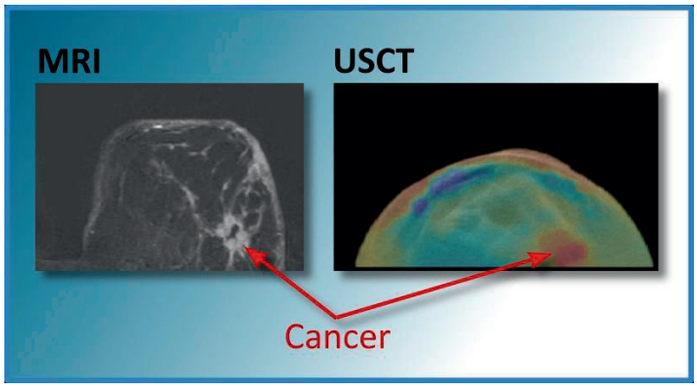


Figure 2: Example of a cancer case with USCT (right) in comparison to the contrast-enhanced MRI image (left).

LITERATURE

[LSDMA] van Wezel, J.; Streit, A.; Jung, C.; Stotzka, R.; Halstenberg, S.; Rigoll, F.; Garcia, A.; Heiss, A.; Schwarz, K.; Gasthuber, M. & others, Data Life Cycle Labs, A New Concept to Support Data-Intensive Science. arXiv preprint arXiv:1212.5596, 2012

[ANKA1] Stotzka, R.; Mexner, W.; dos Santos Rolo, T.; Pasic, H.; van Wezel, J.; Hartmann, V.; Jejkal, T.; Garcia, A.; Haas, D.; Streit, A., Large Scale Data Facility for Data Intensive Synchrotron Beamlines, Proceedings of 13th International Conference on Accelarator and Large Experimental Physics Control Systems (ICALEPS 2011), 2011, pp. 1216-1219

[ANKA2] Yang, X.; Jejkal, T.; Pasic, H.; Stotzka, R.; Streit, A.; van Wezel, J. & dos Santos Rolo, T., Data Intensive Computing of X-Ray Computed Tomography, 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, 2013, 86-93

[UFO1] Vogelgesang, M.; Chilingaryan, S.; Kopmann, A. & others UFO: A Scalable GPU-based Image Processing Framework for On-line Monitoring High Performance, Computing and Communication 2012, IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012, 824-829

[NANO1] Müller, P.; Weiland, Y.; Kaufmann, R.; Gunkel, M.; Hillebrandt, S.; Cremer, C. & Hausmann, M., Analysis of fluorescent nanostructures in biological systems by means of Spectral Position Determination Microscopy (SPDM), In: “Current microscopy contributions to advances in science and technology” (Méndez-Vilas A, ed.), 2012, 1, 3 – 12

[NANO2] Cremer, C.; Kaufmann, R.; Gunkel, M.; Pres, S.; Weiland, Y.; Müller, P.; Ruckelshausen, T.; Lemmer, P.; Geiger, F.; Degenhard, S. & others, Superresolution imaging of biological nanostructures by spectral precision distance microscopy Biotechnology Journal, Wiley Online Library, 2011, 6, 1037-1051

[NANO3] Bohn, M.; Diesinger, P.; Kaufmann, R.; Weiland, Y.; Müller, P.; Gunkel, M.; Von Ketteler, A.; Lemmer, P.; Hausmann, M.; Heermann, D. W. & others, Localization microscopy reveals expression-dependent parameters of chromatin nanostructure, Biophysical Journal, Elsevier, 2010, 99, 1358-1367

[NANO4] Kaufmann, R.; Müller, P.; Hildenbrand, G.; Hausmann, M. & Cremer, C., Analysis of Her2/neu membrane protein clusters in different types of breast cancer cells using localization microscopy, Journal of Microscopy, Wiley Online Library, 2011, 242, 46-54

[LSM1] Mikut, R.; Geurts, P.; Hamprecht, F.; Kausler, B.X.; Marée, R.; Mikula, K.; Pantazis, P.; Ronneberger, O.; Stotzka, R.; Strähle, U. & Peyriéras, N., Automated processing of zebrafish imaging data - a survey, Zebrafish, 2013

[BIO1] Collinet, C.; Stöter, M.; Bradshaw, C. R.; Samusik, N.; Rink, J. C.; Kenski, D.; Habermann, B.; Buchholz, F.; Henschel, R.; Mueller, M. S. & others, Systems survey of endocytosis by multiparametric image analysis, Nature, Nature Publishing Group, 2010, 464, 243-249

[BIO2] Grunzke, R.; Müller-Pfefferkorn, R.; Markwardt, U. & Müller, M., Advancing Cutting-Edge Biological Research with a High-Throughput UNICORE Workflow, Schriften des Forschungszentrums Jülich IAS Series Volume 9, 2011, 35

[DARIAH1] Tonne, D.; Rybicki, J.; Funk, S. & Gietz, P., Access to the DARIAH Bit Preservation Service for Humanities Research Data, 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, 2013, 9-15

[eCOD1] Tonne, D.; Stotzka, R.; Jejkal, T.; Hartmann, V.; Pasic, H.; Rapp, A.; Vanscheidt, P.; Neumair, B.; Streit, A.; Garcia, A.; Kurzawe, D.; Kalman, T.; Sanchez Bribian, B. & Rybicki, J. Stotzka, R.; Schiffers, M. & Cotronis, Y. (Eds.), A Federated Data Zone for the Arts and Humanities, Proc. of the 20th Internat. Euromicro Conf. on Parallel, Distributed, and Network-Based Processing, 2012, 189 - 207

[USCT1] Ruiter, N. V.; Göbel, G.; Berger, L.; Zapf, M. & Gemmeke, H., Realization of an optimized 3D USCT, Proc. SPIE, 2011, 7968, 796805

CONTACTS

[KIT, IPE] **Karlsruhe Institute of Technology, Institute for Data Processing and Electronics**

Swati Chandna, swati.chandna@kit.edu
Robin Dapp, robin.dapp@kit.edu
Hartmut Gemmeke, hartmut.gemmeke@kit.edu
Volker Hartmann, volker.hartmann@kit.edu
Torsten Hopp, torsten.hopp@kit.edu
Thomas Jejkal, thomas.jejkal@kit.edu
Andreas Kopmann, andreas.kopmann@kit.edu
Lukas Niedermaier, lukas.niedermaier@kit.edu
Halil Pasic, halil.pasic@kit.edu
Francesca Rindone, francesca.rindone@kit.edu
Nicole Ruiter, nicole.ruiter@kit.edu
Rainer Stotzka, rainer.stotzka@kit.edu
Danah Tonne, danah.tonne@kit.edu
Xiaoli Yang, xiaoli.yang@partner.kit.edu
Michael Zapf, michael.zapf@kit.edu

[KIT, SCC] **Karlsruhe Institute of Technology, Steinbuch Centre for Computing**

Marcus Hardt, marcus.hardt@kit.edu
Christopher Jung, christopher.jung@kit.edu
Achim Streit, achim.streit@kit.edu
Jos van Wezel, jos.vanwezel@kit.edu

[KIT, IPS] **Karlsruhe Institute of Technology, Institute for Photon Science and Synchrotron Radiation**

Ralf Hofmann, ralf.hofmann2@kit.edu
Julian Moosmann, julian.moosmann@kit.edu
Thomas van de Kamp, thomas.vandekamp@kit.edu
Tomy dos Santos Rolo, tomy.rolo@kit.edu

[KIT, ANKA] **Karlsruhe Institute of Technology, Synchrotron Radiation Facility**

David Haas, david.haas@kit.edu
Wolfgang Mexner, wolfgang.mexner@kit.edu

[KIT, APH] **Karlsruhe Institute of Technology, Institute of Applied Physics**

Andrey Kobitskiy, andrey.kobitskiy@kit.edu
G. Ulrich Nienhaus, ulrich.nienhaus@kit.edu

[KIT, ITG] **Karlsruhe Institute of Technology, Institute of Toxicology and Genetics**

Jens C. Otte, jens.otte@kit.edu
Masanari Takamiya, masanari.takamiya@kit.edu
Uwe Strähle, uwe.straehle@kit.edu

[KIT, IAI] **Karlsruhe Institute of Technology, Institute for Applied Computer Science**

Ralf Mikut, ralf.mikut@kit.edu
Johannes Stegmaier, johannes.stegmaier@kit.edu

[Uni-HD, KIP] **University of Heidelberg, Kirchhoff-Institute for Physics**

Michael Hausmann, hausmann@kip.uni-heidelberg.de
Nick Kepper, nikolaus.kepper@kip.uni-heidelberg.de

[Uni-HD, ERO] **University of Heidelberg, Experimental Radiation Oncology**

Jürgen Hesser, juergen.hesser@medma.uni-heidelberg.de

[Uni-Trier] **University of Trier**

Claudine Moulin, moulin@uni-trier.de
Hannah Busch, buschh@uni-trier.de
Philipp Vanscheidt, pvanscheidt@uni-trier.de

[TU-DA] **Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft**

Andrea Rapp, rapp@linglit.tu-darmstadt.de
Celia Krause, krause@linglit.tu-darmstadt.de

[TUD, ZIH] **Technische Universität Dresden, Center for Information Services and High Performance Computing**

Richard Grunzke, richard.grunzke@tu-dresden.de
Ralph Müller-Pfefferkorn, ralph.mueller-pfefferkorn@tu-dresden.de

LSDMA