ISBN: 978-3-95450-200-4 doi:10.18429/JACoW-PCaPAC2018-FRCA1

NEW COLLABORATIVE APPROACH TO SCIENTIFIC DATA MANAGEMENT WITH NOVA

W. Mexner*, E. Bründermann, M. Caselle, S. Funkner, A. Kopmann,
G. Niehues, N. Tan Jerome, M. Vogelgesang
Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract

Accelerator physics studies at the storage ring KARA at KIT produce terabytes of diagnostics data per day, which is recorded once and then reused on a long-term basis to answer different research questions at KIT. Finally, raw data and intermediate analysis results should be published along with scientific results. Thus storing from the very beginning the data of all analysis steps and its metadata in a central portal would be very beneficial. Similar requirements exist for synchrotron X-ray micro tomography at the KIT imaging cluster and there is an interest to share the large data analysis effort. By using a new collaborative approach, the NOVA project aims to create tools, to enable an efficient use of valuable beam time. For micro tomography beamlines the project will build up a comprehensive database of various demonstrator organisms for the morphological analysis of animals. The NOVA portal is integrated in the local data handling procedures and the datasets automatically appear in the NOVA portal as they are recorded. For both applications, accelerator diagnostics and X-ray tomography, the NOVA portal will offer new collaborative tools to enable synergetic data analysis.

INTRODUCTION

Scientific data management is becoming increasingly important for large scale physics facilities. The European Commission points out already in 2016 that e-science is essential to meet the challenges of the 21st century in scientific discovery and learning [1]. The emerging question is: Is your data useful for somebody in 10 years? Experiments nowadays produce data at extremely high rates, which are put high demands to the whole data acquisition chain with respect to data analysis, curation, storage and usages. While the first two steps are naturally in the focus of the scientists, the later steps are often not handled with the same effort. However, the basics of the data lifecycle must be considered as early as possible so that curated data sets with high-quality content are stored. The experimental boundary conditions must be described in form of metadata as completely as possible in order to be able to analyze data later and reuse it interdisciplinary. For providing open access to these data, the FAIR data principles, as introduced by Wilkinson et al. [2], are important. Data have to be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. To be findable, data should have a globally unique and eternally persistent identifier. To be accessible, also (meta)data have to be online. For being accessible, (meta)data have to be retrievable by their identifier using a standardized open, free and platform independent protocol, e.g. a public repository. To be interoperable, (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. To be reusable, metadata have a plurality of accurate and relevant attributes and should be published together with a clear and accessible data usage license.

Detectors with high temporal and spatial resolution are nowadays available and used in an increasing number of experiments. The resulting data rates are faster growing than data handling technologies. Thus, we move to an era where data becomes to large to copy and more and more scientists in smaller organizations are excluded from latest technologies. During the last ten years several large effort has been taken to find solutions for improved analysis, management and access to large data sets. Synchrotron X-ray tomography has served as an example for imaging applications in general. The integration of online data processing in DAQ systems is essential for high data rate applications. It improves the quality of recorded data and enables advanced experimental control. For the demonstrator application X-ray tomography, a suite of modular software components, called the UFO GPU computing platform, has been designed and implemented. It is intended to execute complex beamline protocols, including real-time data investigation and on-site data processing [3–6]. New methods are added to enhance reconstruction quality and to compute acceptable reconstructions from few data [7, 8]. To manage scientific data at all processing levels, we started the NOVA project. In NOVA, the Network for Online Visualization and Synergistic Analysis, a group of X-ray experts, engineers, computer scientists, mathematicians and biologists teams up to advance analysis tools for tomographic data. The project aims for synergistic data analysis and is building up a comprehensive data portal for morphologic images of small insects [9, 10].

The instrumentation in accelerator physics is currently undergoing a dramatic change. With the new instruments KAPTURE and KALYPSO developed at KIT continuous monitoring of electron beams is possible [11]. These tools uncover phenomena in elctron bunch dynamics and enable advanced beam control [12]. Similar to the imaging applications mentioned before multidimensional datasets with sizes up to the Terabyte level are recorded. Common to all this high rate and high resolution measurements is the need to describe datasets carefully and to provide hierarchical exploration of these datasets. Due to the size of the datasets, storage is costly and access is complex and time consuming.

* wolfgang.mexner@kit.edu

ISBN: 978-3-95450-200-4 doi:10.18429/JACoW-PCaPAC2018-FRCA1

Repositories like the European projects EUDAT or ZENODO allow to store datasets with searchable metadata, but currently there are no suitable tools available to browse this kind of datasets and to inspect multi-scale phenomena. In order to share analysis effort in international communities and to provide access to raw data with scientific publications we will adapt the technologies developed with the NOVA web portal and investigate methods to browse in multidimensional datasets. We are convinced, complex datasets need precise textual metadata and visual exploration. Improved visualization is the key to the development of better metadata for multidimensional datasets.

As a first prototype of the NOVA portal for the tomographic datasets has been already developed. In the next step we will adapt the NOVA environment for electron beam diagnostics at KIT. It is intended to explore with a second community the benefits of collaborative data analysis. The system is constantly evolving and integrated with both user-visible front-end components to enhance the view on the data as well as back-end components that connect the system with the actual experiment setup. In the future, the system will enable scientists to structure their data automatically in a hierarchical manner, which is a prerequisite for efficient storage of large data volumes.

NOVA DATA PORTAL

Synchrotron X-ray microtomography offers unique opportunities for the morphological analysis of small animals. Internal structures become observable even in opaque organisms in a non-invasive, three-dimensional way at sub-micron resolution. By using a new collaborative approach, NOVA aims to create new possibilities allowing for a more efficient use of valuable beam time at tomographic synchrotron beamlines. At the same time the project establishes a comprehensive database of various demonstrator organisms and develops the NOVA data portal to archive, analyze, and share these datasets.

The NOVA data portal is split into a core system and third-party services. The core system provides login, authentication, and management facilities as well as a REST API to interact with those facilities remotely. Third-party services can register with the core system to provide additional functionality and features that should or could not run on the same machine as the portal backend itself. Two main services that are currently implemented are the thumbnail service and a 3D visualization service. The thumbnail service is used to generate iconic images to get a quick idea what kind of data is contained in a data set. The service reads the middle slice of a dataset, shrinks it to a user-defined size and maps the grey values to an RGB triple and saves the resulting JPEG in an on-disk cache from which a browser request is served. Instead of using a generic icon, part of the data with a distinguishing color helps to identify a previously seen data quicker (Fig. 1, left). The 3D visualization service generates on-demand slice maps from the raw image data for client-side rendering using a 3D WebGL browser library WAVe. The service receives a description of the desired dataset and its volume origin and region, then reads the required slices, rescales them and re-orders them in a 2D slice map. This process happens in asynchronous fashion, i.e. the client requests the generation and waits for the service to finish processing and get the final slice map URLs. The slice map's filename is determined by a hash computed from the size requirements for efficient on-disk storage and retrieval. On the front-end side the 3D visualization is started as soon as the slice maps for the entire volume is generated. The user can then translate, rotate, and scale the volume and define a bounding box to "zoom in" to load a smaller part of the volume with a higher resolution. This zoom request will then generate a new task for the slice map server with a new origin and a new region. The right-hand side of Fig. 1 shows the front-end after zooming in on a particular area of the volume.

The WAVe Library for 3D Web Visualization

The WAVe library is a web-based 3D visualization application that renders volumetric slice map data as 3D objects. It consists of multiple modules which render the data in different forms, i.e. using surface rendering or direct volume rendering. These modules are implemented as shaders that enable a high flexibility in fine-tuning the GPU usage such as varying the ray-casting steps or performing various interpolation schemes. The library is based on the Three.js framework which simplifies the 3D scene management. Figure 2 shows the architecture of the library. The API layer allows users to integrate the library into any user interface designs.

ADAPTING NOVA FOR ACCELERATOR RESEARCH

The presented data management workflow has been developed from the beginning in a modular and portable way. It is meant as a generic approach for applications with large multi-dimensional datasets. While being developed for organizing tomographic datasets of biological origin, it can easily being adopted to other fields. We have identified a second use-case for accelerator physics data recorded at the KARA light source, which was also used to acquire tomographic data in the ASTOR campaigns. The accelerator research data e.g. Terahertz research, consists of one Terabyte per machine physics day, with the capability of one Terabyte per hour, and has been acquired by different detectors like EO-Laser, Streak camera, HEB detector, KAPTURE and KALYPSO. The development of fast detection methods for comprehensive monitoring of electron bunches is a prerequisite to gain comprehensive control over the synchrontron emission in storage rings with their MHz repetition rate. For example the "KArlsruhe Linear arraY detector for MHz-rePetition rate SpectrOscopy" (KALYPSO) developed at KIT allows to detect longitudinal electron bunch profiles via single-shot,near-field electro-optical sampling at the Karlsruhe Research Accelerator (KARA) [12]. For one analysis,

ISBN: 978-3-95450-200-4 doi:10.18429/JACoW-PCaPAC2018-FRCA1

100.000 samples with 0,91 Mhz are taken with signal and another set of 100.000 samples as calibration measurements and ar stored structured as tables in the container format HDF5. The organization of the data takes place on file system level with a structuring by date / filling number / detector and measuring station. The metadata is recorded separately in a digital log book (elog) and the accelerator's operating data is automatically logged into a NoSQL Cassandra database.

The interesting use case for NOVA is, that these data has been recorded only once and can be reused on a long-term base for different accelerator physic research questions. Informations like the shift of the bunch profile center in relation to different machine parameters could be automatically calculated and displayed in the portal. Thus, offering all these data enriched with metadata in a central portal would be a big benefit for accelerator research. The core infrastructure is capable of hosting this data, the question that arose was how to ingest it into the system. For the already stored data, some metadata is partially digitally no available (e.g. paper logbooks) or differently structured depending on each experiment type. This means importing can only happen case-by-case. To alleviate this, we identified common metadata structures and began writing an import description format and an import tool that is capable of transforming input data to a common format usable by the NOVA portal. This tool processes an easy extensible job list in JSON format (Fig. 3), which collects all available metadata of the datasets of a certain experiment type and could be added by specific filters collecting data from additional sources like e-logbooks and the Cassandra database.

SUMMARY AND OUTLOOK

Existing repositories like EUDAT or Zenodo allow one to publish digital data with a digital object identifier and make it searchable. The NOVA portal is not only a repository, but allows to browse digital 3D datasets. We are planning to adapt the NOVA collaborative research approach to the

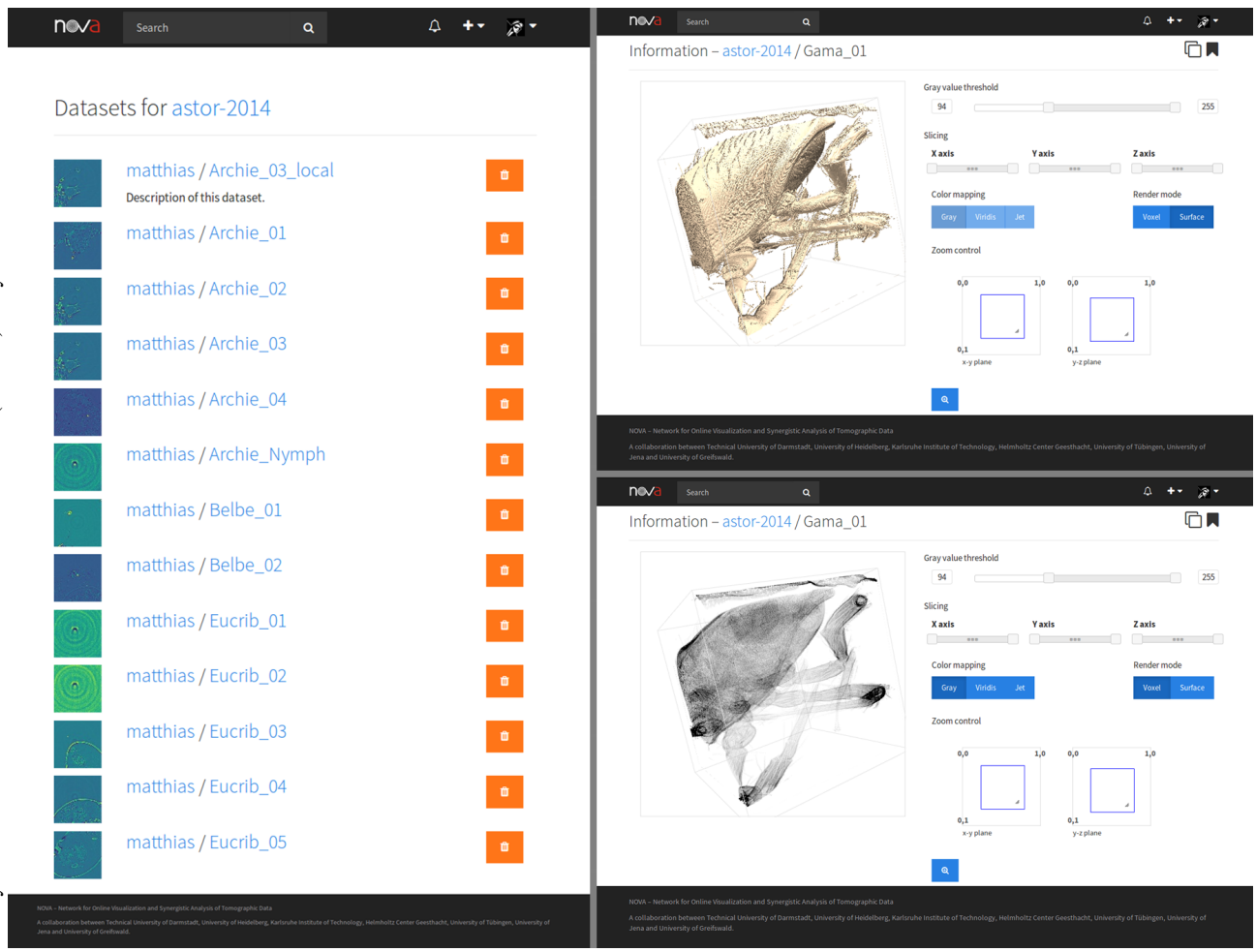

Figure 1: Overview of the NOVA web portal. Left. Thumbnails shown are generated by the thumbnail service allow for easier and quicker distinction between different datasets. Right. Interactive 3D visualization for a selected dataset as surface (upper image) and voxel rendering (lower image). The two smaller blue boxes denote the current zoom area in the x-y- and y-z plane, the blue bounding box in the 3D view shows the current slicing operation.Scheme of the FLUTE accelerator with all installed and planned components.

ISBN: 978-3-95450-200-4 doi:10.18429/JACoW-PCaPAC2018-FRCA1

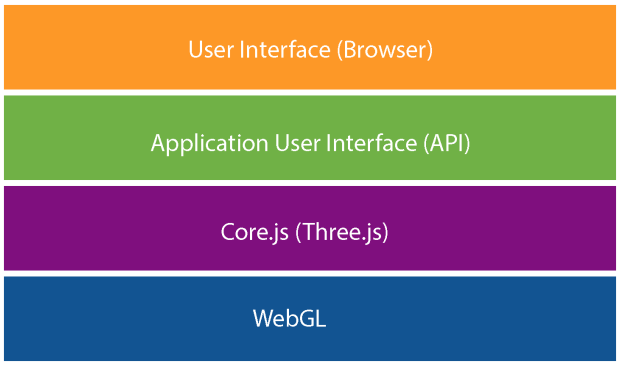

Figure 2: The image represents the high-level architecture of the WAVe library.

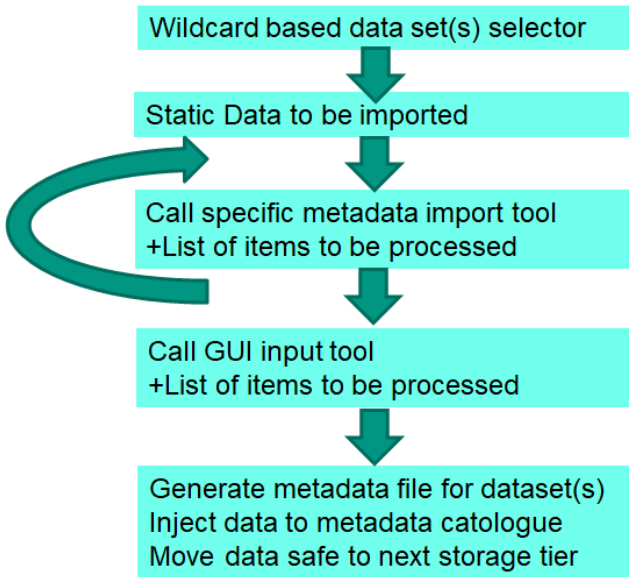

Figure 3: Concept of importing metadata to NOVA portal.

accelerator community for the new generation of detectors for accelerator beam diagnostics.

ACKNOWLEDGMENT

We like to thank the german Federal Ministry of Education and Research (BMBF) supporting the development of the NOVA portal with grant 05K16VBK.

REFERENCES

[1] https://www.naturvardsverket.se/upload/kalendarium/2016/open-data/jiri-pilar-eu-commission.pdf

[2] M. D. Wilkinson, M. Dumontier, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, vol. 3, pp. 160018, 2016. doi:10.1038/sdata.2016.18

[3] M. Vogelgesang, T. Farago, T. dos Santos Rolo, A. Kopmann, and T. Baumbach, "When hardware and software work in concert", in *Proc. IPACLEPCS'13*, San Francisco, CA, USA, Oct. 2013, paper TUPPC044, pp. 661–664.

[4] T. Dritschler, S. Chilingaryan, T. Farago, A. Kopmann, and M. Vogelgesang, "InfiniBand interconnects for high-throughput data acquisition in a TANGO environment", in *Proc. PCaPAC'14*, Karlsruhe, Germany, paper FPO001, pp. 161–163.

[5] M. Vogelgesang, T. Farago, T. F. Morgeneyer, L. Helfen, T. dos Santos Rolo, *et al.*, "Real-time image-content-based beamline control for smart 4D X-ray imaging", *Journal of Synchrotron Radiation*, vol. 23, no. 5, pp. 1254–1263, 2016. doi:10.1107/S1600577516010195

[6] M. Vogelgesang, L. Rota, L. E. A. Perez, M. Caselle, S. Chilingaryan, *et al.*, "High-throughput data acquisition and processing for real-time x-ray imaging", in *Proc. SPIE*, San Diego, CA, USA, 2016, vol. 9967. doi.org/10.1117/12.2237611

[7] A. Shkarin, E. Ametova, S. Chilingaryan, T. Dritschler, A. Kopmann, *et al.*, "An Open Source GPU Accelerated Framework for Flexible Algebraic Reconstruction at Synchrotron Light Sources", *Fundamenta Informaticae*, vol. 141, no. 2–3, pp. 259–274, 2015. doi:10.3233/FI-2015-1275

[8] R. Shkarin, E. Ametova, S. Chilingaryan, T. Dritschler, A. Kopmann, *et al.*, "GPU-optimized direct Fourier method for on-line tomography", *Fundamenta Informaticae*, vol. 141, no. 2–3, pp. 245–258, 2015. doi:10.3233/FI-2015-1274

[9] S. Schmelzle, M. Heethoff, V. Heuveline, P. Loesel, J. Becker, *et al.*, "The NOVA project: maximizing beam time efficiency through synergistic analyses of SRμ CT data", in *SPIE 2017*, San Diego, CA, USA, 2017.

[10] NOVA web page, http://ufo.kit.edu/dis/project/nova.php

[11] M. Caselle *et al.*, "A high-speed DAQ framework for future high-level trigger and event building clusters", *JINST*, vol. 12, pp. C03015, 2017. http://iopscience.iop.org/article/10.1088/1748-0221/12/03/C03015

[12] S. Funkner *et al.*, "High throughput data streaming of individual longitudinal electron bunch profiles in a storage ring with single-shot electro-optical sampling", arXiv.org, arXiv:1809.07530v1 [physics.acc-ph], 2018.