IMPLEMENTATION OF AN OVERALL DATA MANAGEMENT AT THE TOMOGRAPHY STATION AT ANKA

D. Haas, W. Mexner, T. Spangenberg KIT, ANKA, Karlsruhe, Germany
H. Pasic, KIT, Institute for Data Processing and Electronics, Karlsruhe, Germany

Abstract

New technologies and research methods increase the complexity of data management at the beamlines of a synchrotron radiation facility. The diverse experimental data such as user and sample information, beamline status and parameters and experimental datasets, has to be interrelated, stored and provided to the user in a convenient way. The implementation of these requirements leads to challenges in fields of data life-cycle, storage, format and flow.

At the tomography station at the ANKA a novel data management system has been introduced, representing a clearly structured and well organized data flow. The first step was to introduce the Experimental Coordination Service ECS, which reorganizes the measurement process and provides automatic linking of meta-, logging- and experimental-data. The huge amount of data, several TByte/week, is stored in NeXus files. These files are subsequently handled regarding storage location and life cycle by the WorkspaceCreator development tool.

In a further step ANKA will introduce the European single sign on system Umbrella and the experimental data catalogue ICAT as planned as the European standard solution in the PaNdata project.

INTRODUCTION

ANKA is a third generation synchrotron light source operated by the Karlsruhe Institute for Technology (KIT). As an user facility for the national and international scientific community it provides light from the hard x-rays to the far infrared. At the moment ANKA operates sixteen beamlines and is constructing three more. These beamlines are providing research techniques from spectroscopy over diffraction to tomography. Due to this a huge amount of user-, meta- end experimental-data is generated. Currently several 100 TB per year are produced. One of the most data intensive beamlines at ANKA is the tomography station of the Topo-Tomo beamline [1, 2].

These different kinds and huge amount of data leads to the demand of a well-structured and designed overall data management handling all the different types of data. The data management has to cover the whole cycle of an experiment. The data life cycle is starting at the point, where the user applies in the experiment proposal system ANNA for beamtime for the experiment and ending with archiving of the produced data in a well-structured form. So the data management has to consider about the dataflow from the beginning, when the proposal is submitted, over the data acquisition at the beamline, up to the data processing and finally the archiving (see Fig.1).

Figure 1: Dataflow of a whole experiment from the proposal submission to the archiving.

Another goal is to achieve an easy and novel approach to the data from the user's point of view. In the past the data was only organized locally and consist of different types of data formats at several storage locations. To structure these data a lot of manual work had to be done. In addition the users had to copy their data to external mass storages to access and evaluate them afterwards at their place of work.

To overcome this situation a novel overall data management was started to introduce at the tomography beamline TopoTomo at ANKA. Therefore the experiment proposal system ANNA was developed to the needs of ANKA and went productive in the year 2013. Furthermore an Experimental Coordination Service ECS [3] is acting as a state machine of an experiment. The service WorkspaceCreator, claiming, allocating and handling all necessary resources, was developed and introduced.

These mentioned software components are explained and described in terms of scope of duties and technical functionality in the following chapters.

The final realization of implementing all parts is expected in the year 2014.

THE EXPERIMENT PROPOSAL SYSTEM ANNA

The first data is generated when the user applies beamtime for an experiment at the ANNA system. The ANNA experimental proposal system was developed by the Softwareschneiderei GmbH on behalf of the institutes Karlsruhe Nano Micro Facility (KNMF) and ANKA of the KIT making the design layout.

ANNA is the follower proposal system of the Scientific Management Information System (SMIS) [4], which was the proposal system at ANKA over years.

ANNA is a JAVA application which uses Grails [5]. Grails is a powerful framework providing a lot of extra functionality for web applications. It provides also an easy access to the used oracle database for storing all the meta-data of the users and proposals.

After years of operation it becomes more and more difficult to maintain and extend the existing proposal system SMIS. This strongly recommended requirement is provided by ANNA.

ISBN 978-3-95450-139-7

Furthermore is ANNA a modern, intuitive, state of the art web application with a well-organized structure. Thereby it provides the users an easy way to apply for beamtime. Caused by the integration for example to the machine database and the guesthouse reservation, the beamline scientists and the user office get a convenient tool to organize and schedule the beamtime to their specific needs.
To handle the incoming proposals of the users, four user-groups are existing: User, Scientist, Peer Review Board and User Office. The workflow of a proposal submission is shown in Fig. 2.

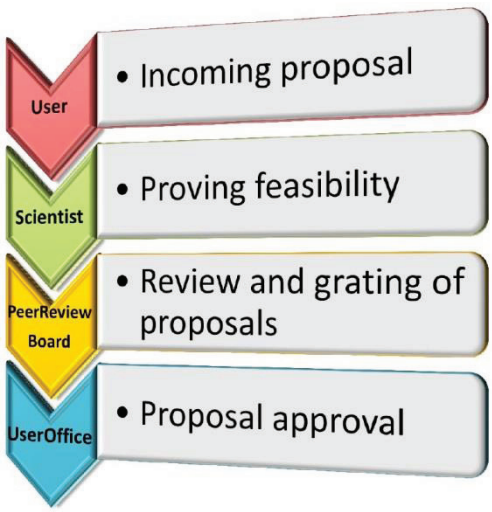

Figure 2: Workflow of applying beamtime. The user submits the proposal. After the feasibility approval of the scientists, the Peer Review Board reviews and grades the proposal. Finally the User Office approves the proposal.

To have an reliable identification, the user gets from the Authentication Authorization Infrastructure (AAI) an unique ID, which ensures the data rights later on.

As a final step ANKA is planning to introduce the European single sign on system Umbrella and an API to the catalogue ICAT of the PaNdata [6] project.

THE RESOURCE MANAGEMENT TOOL WORKSPACECREATOR

Once the experiment is scheduled, resources for the data are allocated by the WorkspaceCreator. The data management is a development of the cooperation of the Institute for Data Processing and Electronics (IPE) and ANKA at KIT.

The WorkspaceCreator is responsible for claiming the necessary resources and initialising the life cycle for each ANKA data set. Each of these data sets is represented in a MySQL database and holds one of the following states, which are necessary to make the data management and data life cycle transparent and consistent:

- '*mutable*' -- The list of dataset components and the corresponding bitstreams may change. Every component has exactly one instance.
- '*immutable-structure*' -- The list of dataset components may not change, but the bitstreams do. Every component has exactly one instance.
- '*immutable*' -- Neither the list of dataset components nor the corresponding bitstreams are allowed to change. A step back to mutable states is possible.
- '*complete*' -- Neither the list of dataset components nor the corresponding bitstreams are allowed to change. A step back to mutable states not possible.
- '*archived*' -- Neither the list of dataset components nor the corresponding bitstreams are allowed to change. A step back to mutable states not possible. The dataset is archived in some repository.
- '*archived-removed*' -- The dataset is archived in some repository, and the data is not present in the experiment domain data management any more.

The previous described status of the datasets can be looked up in a convenient web interface.

This C++ written tool embedded into the Tango environment [7] is collecting the meta-data of the user like name, experiment type, beamtime, beamline etc from ANNA. At the moment the tool is commissioned at ANKA.

Out of these obtained information the WorkspaceCreator is generating a workspace for the user and the experiment. A well organized and reasonable data structure is at the data storage created. This structure contains the user, the name, the beamline and the experiment.

Once the experiment is finished and all data is written into the designated workspace, it is closed and cannot be modified anymore. Closing the workspace ensures the consistence of the data.

THE EXPERIMENT STATE MACHINE ECS

Upon allocating all necessary resources for storing and handling the data, also the retrieval of them has to be organized in consistent and comprehensible way. The ECS is a Tango-server written in C++. It is responsible for managing and coordinating the experimental measurement process. It can be seen as a state machine of the specific experiment.

It is collecting the experimental demands (ED-tags) of the user send through Tango and output a XML file. This queue of ED-tags, defined by the user and written in the XML telegram, is taken down in an ordered queue (FIFO) (see Fig. 4). The demands can consist for example of motor movements or triggering counters. All these movements and triggers are finally executed by the associated Tango device server.

ISBN 978-3-95450-139-7

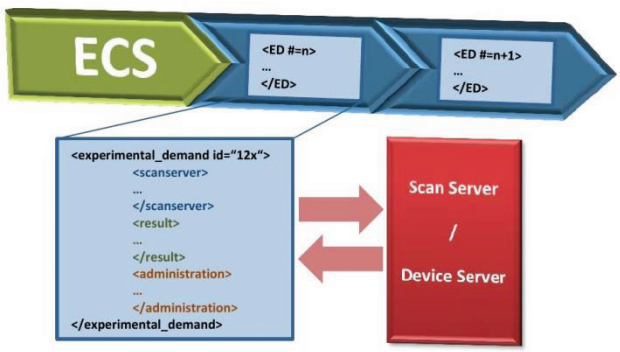

Figure 3: Schematic concept of the Experimental Coordination Service [3].

Finally the ECS is organizing storing the experimental-data together with the meta-data to the preassigned storage location of the WorkspaceCreator.

SUMMARY OF THE WHOLE DATA MANAGEMENT CONCEPT

At ANKA different software components control the dataflow (see Fig. 4). Most of them are connected via ANKA's Tango based environment. The whole concept is designed to contribute a convenient data handling for an experiment at the Topo-Tomo beamline. The dataflow is regulated during registering, performing and storing an experiment.

Starting at the registration, the user is applying for beamtime at the web portal ANNA by a peer reviewed process. The most of the user's meta-data is generated here. This includes for example name, type and time of the experiment. In future it is planned to connect this web portal directly to the Umbrella single sign on system of the European PaNdata project.

With the beginning of the beamtime at the TopoTomo beamline the user authenticates itself with the unique ID into at the AAI into the local control system. By logging in, the user gets a unique, specified workspace, which defines the storage location of the experiment's data. This specified workspace is created and handled by the WorkspaceCreator. The WorkspaceCreator also provides the meta-data of the user. At the moment the meta-data of the user has to be entered manually via a GUI, but the ANNA system containing all these data will be connected soon.

The experiment itself is controlled by the Tango client SPEC [8] and it's easy to handle macro-language. In the case of an ultra-fast tomography it is controlled by the UFO framework [9] and Concert [10], a framework developed in python by the IPE and ANKA [11].

SPEC itself is connected by Tango to the ECS. The Tango-server/client ECS is coordinating the flow of different experimental scans at x-ray beamlines. It gets the demands of a scan from SPEC and executes the tasks in an ordered queue. The ECS is also responsible to collect other experimental data from systems like SCADA WinCC OA [12]. WinCC OA is also visible as a Tango server and client [13]. This data is for example consisting of vacuum pressure and air humidity. After the scan queue is finalized the ECS is saving the meta- and experimental-data in structured way at the given storage location of the WorkspaceCreator.

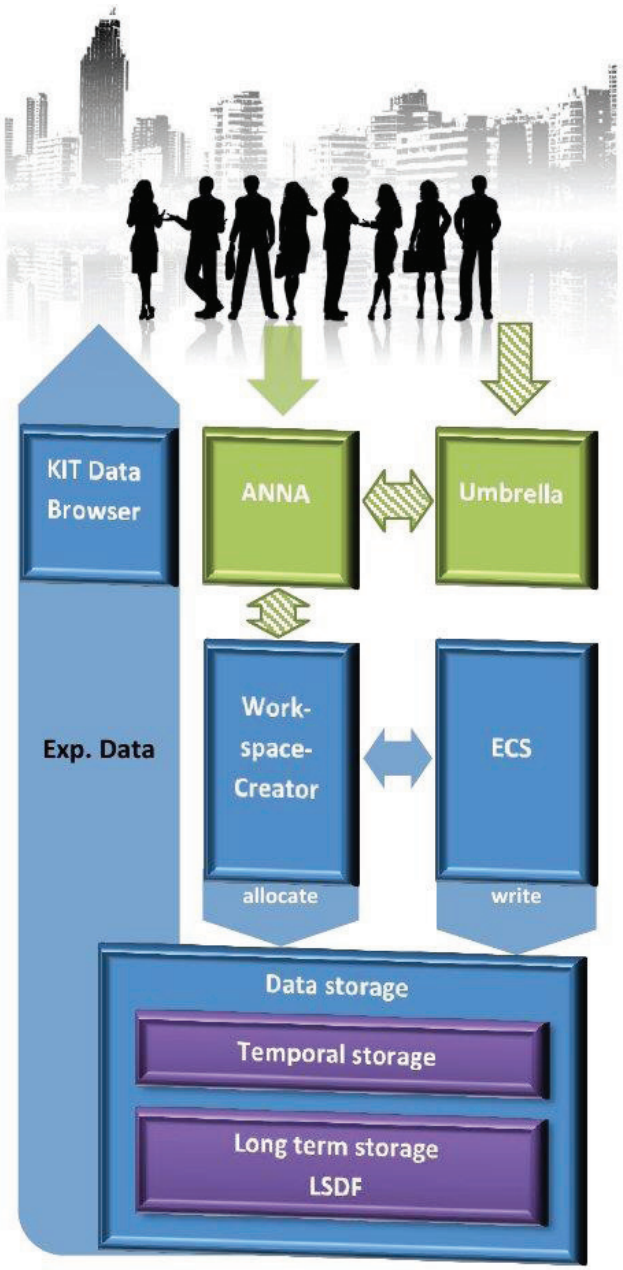

Figure 4: Schematic concept of the dataflow between the different parts of the Tango based control system (the dashed parts are in development).

Like the storage location also the data life cycle is managed by the WorkspaceCreator. After the data is stored at fast local temporal storage at the beamline, the data is automatically shifted to the Large Scale Data facility LSDF [14], a large storage with a volume of several PByte.

Finally the KIT Data Manager provides the data via the KIT Data Browser, which has a web interface for a convenient access to the database, to the users.

ISBN 978-3-95450-139-7

With the final concept the European data catalogue ICAT should be also interfaced.

OUTLOOK

Remaining challenges are to introduce the Umbrella single sign on system and the data catalogue ICAT, which are appropriate to the European PaNdata project. Currently ANKA is working on the full integration of all components in proper and stable way.

ACKNOWLEDGEMENTS

This work has been supported by the Softwareschneiderei GmbH with strong contributions in the development of the ANNA system. Also the Helmholtz Portfolio Extension "Large Scale Data Facility" is acknowledged.

REFERENCES

[1] A. Rack, T. Weitkamp, S. Bauer Trabelsi, P. Modregger, A. Cecilia, T. dos Santos Rolo, T. Rack, D. Haas, R. Simon, R. Heldele, et al., "The micro-imaging station of the topotomo beamline at the ANKA synchrotron light source", *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 267(2011): p. 1978 - 2009.

[2] P. Vagovic, T. Farago, T. dos Santos Rolo, S. Chilingaryan, T. Müller, A. Cecilia, T. van de Kamp, E. Hamann, A. Riedel, D. Haas, W. Mexner, M. Fiederle, T. Baumbach, "Recent upgrades at the micro-tomography station at TopoTomo Beamline at ANKA light source", *will be published in Journal of Synchrotron Radiation.*

[3] T. Spangenberg, D. Haas, W. Mexner, "Smart structured measurement process for versatile synchrotron beamline data at ANKA", *9th International Workshop On Personal Computers And Particle Accelerator Controls*, (2012): Proceedings, p. 40 – 42.

[4] SMIS, http://www.esrf.eu/Infrastructure/Computing/SMIS/.

[5] Grails, http://www.grails.org/.

[6] PaNdata, http://pan-data.eu/.

[7] J.-M. Chaize, A. Goetz, W.-D. Klotz, J. Meyer, M. Perez, E. Taurel, "TANGO – an object oriented control system based on corba", *International Conference on Accelerator and Large Experimental Physics Control Systems*, (1999): p. 475 – 479.

[8] SPEC, http://www.certif.com/.

[9] M. Vogelgesang, S. Chinlingaryan, T. dos Santos Rolo, A. Kopmann, "UFO: A Scalable GPU-based Image Processing Framework for On-line Monitoring", *The Third International Workshop on Frontier of GPU Computing, Liverpool, UK, 25 - 27 June, 2012*, FGC 2012.

[10] M. Vogelgesang, T. Farago, T. dos Santos Rolo, A. Kopmann, T. Baumbach, "When Hardware and Software Work in Concert", *will be published in Proceedings of the 14th International Conference on Accelerator and Large Experiment Physics Control Systems*, (2013)

[11] D. Haas, W. Mexner, T. Spangenberg, A. Cecilia, P. Vagovic, A. Kopmann, M. Balzer, M. Vogelgesang, H. Pasic, S. Chilingaryan, "Status of the Ultra Fast Tomography Experiments Control at ANKA", *Proceedings of the PCaPAC*, (2012): p. 103-105.

[12] WinCC OA, http://www.etm.at/.

[13] T. Spangenberg, K. H. Cerff, W. Mexner, V. Kaiser, "Tango Integration of a SIMANTEC Win CC Open Architecture SCADA System at ANKA", *Proceedings of the 13th International Conference on Accelerator and Large Experimental Physics Control Systems*, (2011): p. 749-752.

[14] R. Stotzka, W. Mexner, T. Dos Santos Rolo, H. Pasic, J. van Wezel, V. Hartmann, T. Jejkel, A. Garcia, D. Haas, A. Streit, "Large Scale Data Facility for Data Intensive Beamlines", *Proceedings of the 13th International Conference on Accelerator and Large Experimental Physics Control Systems*, (2011): p. 1216 - 1219.

ISBN 978-3-95450-139-7