# Trustworthy Laboratory Automation

Jan Potthoff, Dominic Lütjohann, Nicole Jung
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
[name.surname]@kit.edu

***Abstract*—To ensure the quality of scientific data, the integrity and authenticity of the data has to be guaranteed. Due to the significance of the primary data, the integrity and authenticity of research data have to be ensured with their generation. Special measurement devices offer this possibility by an automated digital signing process. Unfortunately, these devices are rare. Therefore, a generic software application which is shown in this paper has been implemented. Furthermore, the solution has been adjusted to an existing laboratory automation system which is depicted as well. The combination of a new, web-based Laboratory Information and Management System (LIMS) with a new protocol for an authentication of electronic data shows that the establishment of an automated collection of secure data can be realized even without disturbing the researcher.**

***Keywords-integrity and authenticity; sustainability; trustworthy data generation; data management***

## I. INTRODUCTION

In the field of natural sciences, data collections and tools for their valuation are of enormous importance as primary research data influence directly all experiment-driven interpretation [1]. Consequently, these primary research data form the basis of common knowledge. Nonetheless, the collection of scientific data sets –especially in academic labs– is very often not an automated process [4]. Up to now, the single researcher is responsible for the selection of the most important data, for the storage of these data and the researcher has to approve the integrity and authenticity of the information [2]. This traditional procedure of data acquisition will probably never be replaced by software-aided processes if the decision of the researcher is necessary.

In contrast, other processes as the acquisition of data generated by half-automated lab devices seem to be suitable for an automated collection, certification and management procedure. Examples of the latter operations are the analytical instruments that run with their own proprietary firmware giving standardized datasets that are the basis of qualitative and quantitative measurements [3]. Data collection of data produced by such devices means on the one hand the implementation of an automated procedure for the permanent actualization of the given research data (for each project, researcher, or group) and on the other hand the implementation of a reliable protocol for a doubtless statement on the integrity and authenticity of the gained datasets. Both challenges have been addressed and are described herein via the presentation of selected examples (Section I to IV).

In Section II, the requirements of analytical devices for long-term preservation and the motivation for developing tools for trustworthy data management are depicted. A detailed example of the data generation process is added via description of an automated data collection process in Section III. Finally, the implementation of a ready to use software tool (Section IV) is exemplarily shown to demonstrate the applicability to gain trustworthy primary data in a lab environment.

## II. SCIENTIFIC DATA LIFECYCLE

In general, the research process is highly individual and differs from one scientific branch to another and from one researcher to another. However, the experimental scientific process flow may be roughly divided into five phases (planning, implementation, analysis, publication and archiving) [5]. In each of these phases, digital data is produced especially by measurement devices and the usage of computers.

### *A. Measurement Devices and Scientific Data*

The equipment of an exemplarily chosen chemical lab consists of three different groups of devices. They are divided into synthetic devices (as shakers, stirrers, heaters, microwaves and reactors), purification stations, e.g., preparative high-performance liquid chromatography (HPLC), medium pressure liquid chromatography (MPLC) and analytical devices as liquid chromatography mass spectrometry (LCMS), gas chromatography mass spectrometry (GCMS), nuclear magnetic resonance (NMR), ultra violet (UV), infra-red (IR), and Raman spectrometers. Whereas the devices in the first group are often standalone devices with almost no functionality for electronic data storage, the devices of group 2 and 3, namely the purification stations and the analytical devices, contain intern intelligence for the acquisition and management of electronic information. Therefore, the investigations on the storage and authentication of scientific data focus initially on the analytical devices which process all available information and which are of highest interest as they deliver the final data of the research projects.

If possible, the data format is chosen by its creator. Because of several data sources, e.g., software programs or measurement devices, as described above, a free choice is not always possible, proprietary data formats have to be

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-334-6 

accepted and the data migration becomes difficult. Based on several catalogues of criteria, the data formats can be evaluated according to their suitability for long-term preservation. It is important to keep in mind that the choice of data format should respect the actual situation but also the archival process [6].

### B. Trustworthy Data Management

Data processing systems have to meet several requirements to ensure the quality and protection of data. For example, the IT-Grundschutz Catalogue of the Federal Office for Information Security (BSI) defines four core requirements namely availability, confidentiality, integrity and authenticity referring to information security [7]. These criteria are very important for all processes that are based on a high level of trustworthiness such as archival systems. In this case, the availability, confidentiality, integrity and authenticity have to be ensured for a long time. Therefore, the reference model Open Archival Information System (OAIS) defines essential functions and processes to guarantee these requirements [8].

As most of the available data will not only be generated and archived but will also be used, the data have to be managed by specific programs or individual folder structures. The type of data management depends on the application, the requirements of the organization and the individual requirements of the user. For example, in research, Electronic Laboratory Notebooks (ELN) or Laboratory and Information Management Systems (LIMS) are used for the data management [9]. These paperless lab management alternatives offer many advantages in contrast to the documentation on paper, but they suffer from one important disadvantage: changes in digitally stored data are not detectable.

Within the DFG project “BeLab” (probative electronic laboratory notebook) an interdisciplinary work group analyzed how accessibility, completeness, integrity, authenticity, readability and interpretability of ELNs can be ensured in the long term [26]. The result of the project is a service (BeLab system) which ensures the integrity and authenticity of the submitted data. As different research areas use individual scientific tools in their processes, the BeLab system has been designed as a generic data verification system which uses multiple ingress, verification and egress modules. By using the service, the provability of scientific data in digital archives is profoundly enhanced [10]. In [11], a generic solution for data management according to the Good Scientific Practice (GSP) is depicted as a further result of the BeLab project. By this application the scientists are able to archive their data with respect to the GSP. The data can be managed by a Graphical User Interface (GUI) which offers the possibility to collect files to be archived or to check out archived files and edit them. Additionally, metadata which refer to the structure of data, e.g., project, ELN and general container id can be added. These metadata can be indicated for each added file in the data section as well. Examples for such additional descriptive information are document title, author and creation date. This generic solution prepares the managed data for a trustworthy archiving. It shows that the integrity can be easily implemented within the data management environment. Regarding the long-term preservation, the data can be submitted to the BeLab system.

### C. Related Work

The integrity and authenticity of scientific data must be ensured in regards to several regulations [12][13]. In addition, all research fields have to guarantee a sustainable archiving of this data [14]. To foster and to assess the trustworthiness of archival systems, further research projects have addressed the issues of availability, confidentiality, integrity and authenticity, as described in the previous section [15][16].

To ensure the integrity and authenticity, the research process flow has to be considered as well. Current available systems are tailored for specific device manufacturers usually commercially available as integrated and monolithic lab automation solutions [17]. Customization and expansion, e.g., the integration of new devices or devices from different vendors, requires drivers and software components specifically developed for the platform and are usually not usable in other environments.

Open Source solutions [18] for lab automation allow flexible implementation of new devices, but are usually not prepared for data acquisition in a dynamic lab infrastructure which requires frequent changes in configuration such as laboratories in academic facilities. Furthermore, installation and maintenance requires investments in server infrastructure, as long as data ownership needs to be considered and commercial cloud services cannot be taken into consideration.

## III. LABORATORY AUTOMATION

The automated data generation process is the key requirement in data-driven laboratories. Existing devices and computer systems need to be integrated into unified views of the overall data structure of a laboratory to allow scheduling of tasks, remote operation of devices, sample tracking, data management, and archiving.

### A. Practical Requirements

Many laboratory devices are equipped with a command and control personal computer system, which is connected via ethernet connection or other serial interfaces to the device. To operate the device, vendor-specific software needs to communicate with the device using device drivers and command sets which often follow proprietary protocols. To integrate these kiosk-like systems in an automated laboratory, an overlay software is required, which does not need deep modifications on operating system and application level. Furthermore, it should not be necessary to adjust the data storage routines and locations, to access the created datasets in a LIMS.

### B. Webbased Device Control

Once samples are loaded onto a device, the data acquisition process takes time, depending on the selected method. To allow users to operate devices remotely, e.g.,

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-334-6 

from different workplaces and different mobile devices, and observe the output values and analysis status, the user interface needs to be shared on a web-based platform. The experimenter who operates the device also needs control over the generated datasets from a centralized platform to integrate the contained results into a research and results database. Beginning from this step, identification of the user is possible and therefore the ownership of datasets can be determined. This allows data policies to be applied at the time of data collection and enables ubiquitous access.

## IV. TRUSTWORTHY DATA COLLECTION

The amount of digital data is constantly increasing. To be able to manage these data, special software and hardware solutions have been developed over the last years. Procedures to ensure the integrity and authenticity of the data in paper documentations have already been established but these regulations have to be transferred to the electronic datasets as well. To ensure the trust in new, automated technologies, specific requirements have been described. The obligation to preserve records also applies for these digital data. This leads to several requirements regarding the digital long-term preservation as well. One of these requirements is the integrity of all produced data: "Integrity considers all possible causes of modification, including software and hardware failure, environmental events, and human intervention." [19]. The combination of these tools will end in a trustworthy archive, which should prevent the quality of data. In addition to that, the quality of data depends on its generation, processing and preparation. To ensure the integrity and quality of all given information, the overall process of their generation has to be analyzed.

### A. Practical and Legal Requirements

The importance of the verification and integrity of data is always desirable and in many cases absolutely mandatory. Due to fast proceedings of computing and digital processes that find their way into almost all areas of life, the legislative organs worldwide have to be concerned with regulations upon these new paperless developments. Whereas, in general, all data can be easily manipulated, digital data are even more prone to undesired alteration as changes –even in much larger scale– can be easily made without being visible to others. In the year 2001, the German law responded to these developments and special bills have been developed for digital data. By these bills, the evidence of digital documentation is regulated based on the digital signature. This signature can be used to prove the integrity of digital data and the authenticity can be shown by the corresponding certificate [20]. A further opportunity to validate the integrity of digital data is the qualified digital time stamp, which is required in several organizations and which can be used to prove the existence of a document. Whereas other alternatives offer very similar opportunities, the processing of a qualified digital stamp is the only procedure that has been accepted by the German law [21].

To ensure a high probative force of digital data, the integrity and authenticity has to be ensured as early as possible, which usually means at the same time as the data is generated [22]. A lot of analytical devices which generate digital data are used for experimental research, so that the best possibility would be to ensure the integrity and authenticity by these devices. This kind of hardware-triggered solution could be implemented as integrated software package within the analytical device. A few of these analytical devices with implemented digital signing procedure are available today, but adopting this solution will result in the dependency on the manufacturer. A generic approach should be addressed giving a solution that should be independent of the used devices, the research area and the used hardware. The only restriction to be made is that it will be assumed that the analytical device is connected to a computer. Due to the independency of the research area, the solution has to be adaptable to further requirements of individual use cases.

Due to the juristic requirements, as previously described, the solution should use digital signatures to ensure the integrity of data. The development of such a signature trail is always the result of a compromise: As the researcher should not be affected by the signing process, the digital signature should be added, on the one hand, in an automatically manner. On the other hand, the automatic signing process decreases the probative value of the signature. Therefore, the process has to be associated with the person [23].

Because a generic approach will be separately implemented from the analytical devices, the time between data generation and integrity protection is of prime importance. The period of time should be as short as possible. In addition, other organizational measures, e.g., room entry controls, can be also implemented to ensure the data integrity. The implemented safety measures should be transparent as well as the functioning of the solution to reach the trust in the solution.

To keep the reached probative value, a trustworthy archive should be used at the end. Because the measurement data should not be altered, the archival process can be directly started after the data generation. By doing this, the data sharing and usage is reached as well.

### B. Protected Data Collector

A secure data generation can be reached by digitally signing analytical devices. These devices have an internal cryptographically chip that calculates the signature. By doing this, the data is signed before leaving the device and a data manipulation can only be reached by a manipulation of the device. By sealing the devices, the manipulation can be proven and the integrity and authenticity can be validated by the signature and the corresponding certificate.

According to the practical and legal requirements, as described in Section IV-A, and to become more generic, a software solution (hereafter called data collector) has been developed to offer the functions of data integrity and authenticity for individual analytical devices. In a first step, a generic interface between device and the data collector, e.g., a folder for the data transfer, has to be defined. Using such a folder, analytical devices store their data via a gateway that can be monitored by the data collector. If the data collector

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-334-6 

detects new files, it will execute all necessary steps to ensure the integrity and authenticity of the processed data.

A particular challenge is the definition of the time of archiving. For this purpose, one folder can be monitored by several individual modules. For example, a module monitors the number of files in the corresponding folder. If a defined number of files are detected, the archival process will be started. These modules can be further specified in the system configuration offering the possibility for combinations of different modules and the integration of individual modules.

If a new file is detected, the corresponding hash value will be calculated and stored in the main memory. From that on, the integrity of the detected file is temporarily secured. After each process, the individual modules check the archiving condition and the results will be logically combined to one outcome. All hash values will be calculated again prior to the archival process and they will be compared with the values in the main memory. If no changes have been detected, the data will be digitally signed by the data collector. Before this, the detected files will be prepared for the long-term preservation which includes the package of the analytical data into a data container, called Universal Object Format (UOF) [24]. This container which includes integrated metadata has been developed by a German research project of long-term preservation [14]. The metadata format is based on the Metadata Encoding and Transmission Standard (METS) [25] that includes attributes of file, e.g., filename and hash value. With this information, the integrity can be proven at any time. Whereas this procedure still allows the manipulation of files and hash values, the metadata file is digitally signed so that all changes can be proven. By doing this, the legal requirements, depicted in Section IV-A, are addressed.

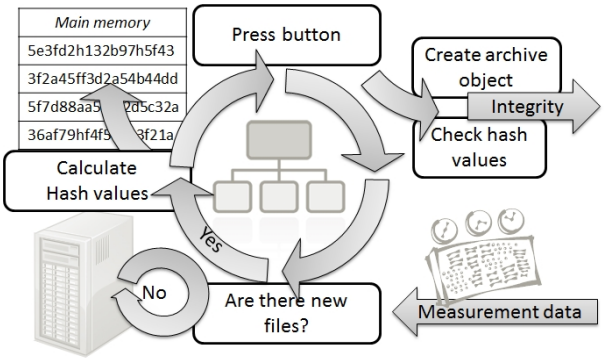


Figure 1. Working process of data collector

After finishing the process mentioned above, the data will be submitted to an archive interface. This interface checks the data according to the requirements for long-term preservation and its probative value. For example, existing digital signatures will be validated again. In addition, a specified module checks the completeness, e.g., filename dependencies are analyzed.

The data collector can be used in the data management process, as described in Section III, if it is adjusted according to the time of archiving. Due to an architecture where an individual amount of files belongs to one output directory and several subfolders belong to one measuring, a corresponding archiving module cannot be implemented for these data management purposes. As a solution, one root folder is monitored in which a new subfolder will signal new experiments that have to be monitored. After this initial procedure, all files and subfolders, which are stored in this folder, will be noticed by the data collector and the integrity is secured as previously described.

The archiving starting point can be signaled by the user with a small GUI. This GUI contains a button which is inactive if no new subfolders are noticed. If a new folder is noticed, the button will become active and the user can start the archival process at any time. The process of securing the integrity and archiving by the data collector is depicted in Figure 1.

By pressing the button, all noticed subfolders of the root folder will be prepared for an archive object and will be submitted to an archive interface as previously described. After that, the button will be disabled again until new subfolders are founded.

Naturally, the researcher can modify the primary analytical data after the data has been submitted to the archival system. For this purpose the archival process has to be adapted. The data collector will notice the changes of the corresponding files. If files, which have been already archived, are changed and the subfolder is noticed again, the archived data must be updated. For this the archival process can be restarted again by the user. To handle all versions of the data, all IDs received from the archival system and the corresponding folder names are stored. If the folder name already exists in this list, the update function will be used in combination with the corresponding ID. This way, the integrity of measurement data is guaranteed and all research results are traceable.

With this solution, the measurement device must not calculate the required digital signature by itself. Thereby the data integrity assurance can also be used for devices with a fast data processing. Additionally, the archival process is separated from the data processing so that the scientist is not affected.

### C. Safety Aspects

The highest probative value is reached by digital measurement devices which offer the possibility of internal digital signatures. On the other hand, this function is not implemented in many devices yet. Therefore, a mechanism, as described in previous Section, is needed which can be used in combination with each measurement device. However, same security gaps have to keep in mind.

The first challenge is the data transfer from the analytical device to the connected computer. During this procedure, manipulations are most likely in particular if the device is connected via network. In this case, a secure data transfer protocol like https has to be used. If the files are stored, the data collector will need a short time until it detects the new files and a security gap results. In consequence, data manipulation may occur between the finishing of the file and its detection. This fact can be addressed technically by

minimizing the time gap. If operating system-based solutions are used, the time gap will reduce to a few milliseconds. By doing this, a data manipulation is improbable.

After the detection of new data the hash values will be calculated by the data collector and stored in the main memory. Theoretically the values in the memory can be manipulated on a higher level. This manipulation would mean a high criminal energy. Generally, the solution is not developed to prevent any faults but rather to give a solution to demonstrate the appropriate work of the researcher. In the final step, a manipulation can be made during the data transfer to the archival system but the used archive interface is based on the https protocol so that a secure data transfer is given.

On the juridical side, the focus is on how probable a manipulation is. If a network is used, the mistrust will be high even if the data transfer is encrypted, whereas a separated room with controlled human access leads to a higher trust. Data processing is not the critical point if a secure environment can be established. Generally, digital data can be manipulated easily. Therefore, particular laws have been established by the German legislation to rule the usage of digital signature and to get a probative force. Therefore, the submitted data will be automatically signed by the data collector. According to the usage of an automatically process for digital signatures, it has to be noted that the procedure has to be connected to a person, which means that a person has to start and initialize the system with its corresponding certificate [23]. Only then a digital signature is comparable with a handwritten signature according to the German law. The same applies to signing measurement devices.

## V. Conclusion and Future Work

The implemented software application data collector aims at providing a generic solution for the guarantee of the data integrity and authenticity. In comparison with digital signing measurement devices some safety risks have to be accepted. However, these risks can be reduced by administrative measures. The software application has been adjusted for the depicted workflow. It can be seen that the adjustment of the archiving condition was difficult because of individual data structures of the lab devices.

The depicted software solution fulfills the requirements of the German electronic signature law and the regulations of the GSP. For general usage, further regulations have to be considered.

The workflow in chemical labs is only one example for the need of professional data handling but forms an ideal model for the implementation of an automated data collector. Most of the processes in chemical labs underlie standardized procedures and the main datasets consist of values that have been acquired from well-defined, standardized instruments. In order to identify these processes in a chemical lab, the dataflow in an organic working group has been investigated. Advanced analytical instruments as NMR-spectrometers, chromatography-mass spectrometers (liquid and gas chromatography, HPLC/GC) and electron ionization (EI)-mass spectrometers were determined as devices with high priority for the definition of experimental conclusions and for the confirmation of the results in publications. The implementation of a data collector covering results that have been processed by these instruments was starting point of a new interdisciplinary project.

## References


[1] J. G. Frey, "Dark Lab or Smart Lab: The Challenges for 21st Century Laboratory Software," Org. Process Res. Dev. 8, 2004, pp. 1024-1035.

[2] C. L. Bird and J. G. Frey, "Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences," Chem. Soc. Rev. 42, 6754.

[3] R. Bramley, K. Chiu, T. Devadithya, N. Gupta, C. Hart, J. C. Huffman, K. Huffman, Y. Ma, and D. F. McMullen, "Instrument Monitoring, Data Sharing, and Archiving Using Common Instrument Middleware Architecture (CIMA)," J. Chem. Inf. Model., 2006, 46 (3), pp 1017–1025.

[4] N. Jung and D. Lütjohann, "Chemie auf der Spitze des Eisbergs: Zu viele Forschungsdaten gehen bislang unter!," Chemie in unserer Zeit, vol. 47, 2013, pp. 334-335, DOI:10.1002/ciuz.201390062

[5] S. Hackel, P.C. Johannes, M. Madiesh, J. Potthoff, and S. Rieger, “Scientific Data Lifecycle – Beweiswerterhaltung und Technologien,” Proc. 12. Deutscher IT-Sicherheitskongress (BSI-IT-SEC 2011), SecuMedia, 2011, pp. 403-418.

[6] A. Brown, “Digital preservation guidance note: Selecting file formats for long-term preservation,” 2006.

[7] Federal Office for Information Security (BSI), "IT-Grundschutz Catalogue," 2005, https://www.bsi.bund.de/EN/Topics/ITGrundschutz/itgrundschutz.html 11.02.2014.

[8] CCSDS, "Reference Model for an Open Archival Information System (OAIS)," CCSDS, 2002.

[9] M. Rubacha, A. K. Rattan, and S. C. Hosselet, “A Review of Electronic Laboratory Notebooks available in the market today,”, JALA, vol. 16, Feb.2011, pp. 90-98, doi:10.1016/j.jala.2009.01.002.

[10] J. Potthoff, S. Rieger, and P. C. Johannes, "Enhancing the Provability in Digital Archives by Using a Verifiable Metadata Analysis Web Service," Proc. 7th ICIW 2012, 2012, pp. 112–117.

[11] J. Potthoff, M. Walk, and S. Rieger, "Data Management According to the Good Scientific Practise," Proc. 5th DBKDA 2013, 2013, pp. 27–32.

[12] Deutsche Forschungsgemeinschaft, "Recommendations of the Commission on Professional Self Regulation in Science - Proposals for Safeguarding Good Scientific Practice," January, 1998, http://www.dfg.de/en/research_funding/legal_conditions/good_scientific_practice/ 26.02.2014.

[13] OECD, "OECD Principles on Good Laboratory Practice," Paris, 1998, http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/mc/chem%2898%2917 26.02.2014.

[14] R. Altenhöner, "Data for the future: The German project ‘Co-operative development of a long-term digital information archive’ (kopal)," Library Hi Tech, Vol. 24 Iss: 4, 2006, pp. 574–582, doi:10.1108/07378830610715437.

[15] Trustworthy Repositories Audit & Certification (TRAC), "Criteria and Checklist, Center for Research Libraries," OCLC Online Computer Library Center, 2007.

[16] N. Beagrie et al., "Trusted Digital Repositories: Attributes and Responsibilities," RLG-OCLC Report, 2002.

[17] Agilent OpenLAB, http://www.chem.agilent.com/en-US/products-services/Software-Informatics/OpenLAB-CDS-Chemstation-Edition/Pages/default.aspx 26.02.2014.


Copyright (c) IARIA, 2014. ISBN: 978-1-61208-334-6 

[18] Bika LIMS, http://www.bikalabs.com 26.02.2014.

[19] T. Malone, G. Blokdijk, and, M. Wedemeyer, "Itil V3 Foundation Complete Certication Kit-Study Guide Book and Online Course," Lulu. com, 2008.

[20] S. Mason (Ed.), "International Electronic Evidence," London, UK: BIICL, 2008.

[21] D. Huhnlein, "How to qualify electronic signatures and time stamps," Public Key Infrastructure, Proceedings, Lecture Notes in Computer Science, 2004, pp. 314-321.

[22] J. Potthoff, S. Rieger, P.C. Johannes, and M. Madiesh, "Elektronisch signierende Endgeräte im Forschungsprozess," Proc. D-A-CH Security 2011, syssec, 2011, pp. 44-55.

[23] A. Roßnagel, S. Fischer-Dieskau, "Automatisiert erzeugte elektronische Signaturen," MMR 2004; S. 133 – 139.

[24] T. Steinke, "The Universal Object Format – An Archiving and Exchange Format for Digital Objects," in Research and Advanced Technology for Digital Libraries, Springer Berlin, 2006, pp. 552–554.

[25] Digital Library Federation, "<METS> Metadata Encoding and Transmission Standard: Primer and Reference Manual," Version 1.6 Revised, 2010, http://www.loc.gov/standards/mets/ 26.02.2014.

[26] BeLab project, http://www.belab-forschung.de 26.02.2014.

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-334-6 