IMPROVING THE TRADITIONAL INFORMATION MANAGEMENT IN NATURAL SCIENCES

Martin Kühne and Andreas W. Liehr*

Freiburg Materials Research Center, University of Freiburg, Stephan-Meier-Str. 21, 79104 Freiburg, Germany
**Email:* liehr@fmf.uni-freiburg.de

ABSTRACT

It is common practice that in teams working in the field of natural sciences all group members manage their primary data in highly individual systems. The consequence of this is that the data are usually lost shortly after the scientific results have been published or that they lose part of their value, as significant background information can no longer be found. To solve this problem in a simple way, we present a basic procedure that allows us to uniquely identify scientific data and their history at any time. We describe which requirements such a procedure has to meet (proper documentation, unique identification, and easy backtracking of the individual operations) and discuss on the basis of a timestamp approach how such a procedure can be integrated smoothly into the traditional scientific work process and the existing IT infrastructure of a team. It does this by using established processes and developing them into a systematic information management covering both electronic and analogue media.

Keywords: Information management, Primary data, Laboratory notebook, Timestamps, Knowledge chains

1 INTRODUCTION

Natural sciences have become increasingly important in modern society to safeguard progress and the standard of living (Mlynek, 2007). This development is associated with rising demands on scientists and engineers. For example, productivity is to be increased while the quality of the results should remain the same or be even higher. Moreover, the pressure to publish research results has increased dramatically in academic fields. In addition, the introduction of computer-assisted, automated, or novel measuring and evaluation techniques has brought about an enormous rise in the quantities of data to be interpreted (Denning, 2007). And finally, as if this were not enough, these data have to be exchanged with specialists from other technical fields because of the increasingly interdisciplinary approach of research. It goes without saying that transparency and verifiability also have to be guaranteed at any time (European Commission, 2005).

Presently, many procedures and information technology (IT) infrastructures designed to solve these problems have been developed, tested, or already used. In this context, approaches such as self-documenting data formats (Zimmermann et al., 2007; Hartnett & Rew, 2008), databases (NIST, 2005; Signalling Gateway, 2008), laboratory information systems (McDowall, 1999), electronic laboratory notebooks (Butler, 2005; Belenkaia et al., 2006), and scientific information repositories (Papiani et al., 1999; Tuchbreitner et al., 2003; Buneman et al., 2004; Riede et al., 2008) for the exchange of scientific primary data are of utmost importance. Irrespective of their differences, the common denominator of all these IT infrastructures is that they are developed for managing huge quantities of information. This may be the reason why the majority of scientific work groups still rely on the traditional way of information management, which is based on the classic laboratory notebook and standard computer file systems.

This traditional information management approach works quite well for the experienced scientist contributing rather independently to the research efforts of his or her work group in physics, chemistry, biology, or even engineering. Transferring this knowledge to coworkers or successors is inefficient, however, due to the fact that each scientist organizes analog and digital information repositories with an individualized systematic order. The resulting inefficiency increases with the number of work group members and the rate of staff fluctuation, which is rather high especially in academic fields. As a consequence, many scientists and Ph.D. students have already been faced with the problem that older measuring data could no longer be interpreted because the complete history of these data could not be reconstructed, e.g., which samples were actually measured, which experimental parameters were applied, or which properties (e.g. the molecular weight and polydispersity of polymer samples) did the tested

substances have. Of course, such problems may also arise due to generally incomplete or incorrect documentation, but this scenario is not discussed in this paper, as such fundamental problems cannot be remedied by simple means.

In most cases, however, the documentation is quite adequate, but the general context is lost because of the multitude of steps to be documented separately and the elusive relationships among individual analog and electronic documents (e.g. protocols in laboratory journals, primary data files on hard disks, etc.). The consequence is that the origin of specific information cannot be fully reconstructed or backtracked any more, thus making them useless for serious scientific work. This is a waste of knowledge and also of effort, as in most cases the measurements have to be repeated. It is therefore important to use a method that prevents such frictional losses. However, this method should preferably not interfere with the working method presently used by the scientists, but be based on traditional structures, as experience has shown that changes requiring an excessive amount of adjustment are not accepted by the scientific community (Hodgson, 2001; Herb, 2008). The objective of this paper is to point out in concrete terms the special situation of the management of knowledge in natural sciences and to describe a general information management system that takes this special situation into account and is based on simple rules.

2 THE FLOW OF INFORMATION IN NATURAL SCIENCES

The workflow of a scientist is usually organized by two processes: the characterization of an object by experimental or analytical methods and the creation of new objects from already existing objects. The special challenge of information management in the field of natural sciences is that the investigations involve both material and immaterial objects to achieve scientific results (Polonsky et al., 2006). Let us illustrate this abstract description with an example from materials science (Figure 1). In this example, two material objects (polymer *A* and polymer *B*) are processed into a new object (sample) in one processing step. This material object is characterized by a measurement creating an immaterial object in the form of a primary data file (result.dat). For the materials scientist, all these objects are information bits of one or another forming the overall picture, which is completed by the metadata on the polymer specifications, the blending process, and the experimental setup. When all these relevant information bits and their appropriate documentation are connected to each other, a chain of research knowledge is formed.

The standard procedure used by work groups engaged in scientific research is that each scientist documents his work individually in a laboratory journal (Ebel et al., 2006) as a set of successive protocols, which usually are annotated by their creation date. Often the individual measuring files are not even stored on a central server but in a distributed data system. In addition, the documents regarding the source materials, e.g. safety data sheets or specifications of the manufacturer, are collected and archived in a central reference library.

In the normal academic daily routine, this creates a multitude of potential sources of error. Often it is difficult - especially if many reiterated measurements are performed on a sample over a prolonged period of time - to link a file of measurements clearly to the relevant measurement protocol. Or even worse, a measurement or primary data file is assigned to a wrong sample because there are several samples of the same name. Usually scientists can keep such kinds of errors in check because they remember and know their personal method of knowledge management. However, in case of doubt, samples have to be prepared again, new measurements have to be carried out, or new simulations have to be performed. This happens frequently to young scientists because at the beginning of their career (e.g. as PhD students), they do not yet have the experience to handle information volumes of larger projects. Thus they are often faced with the problem at the end of their doctoral thesis that the results of the first year can no longer be fully used, simply because in the course of their doctorate, they changed their management approach several times. In addition, serious problems may occur when a scientist leaves an institute and other researchers have to use his or her knowledge chains compiled in the form of samples, laboratory journals, and primary data files.

Especially in the academic field, this situation is the rule rather than the exception. Nevertheless, it should be possible without any problems to continue using the data of the predecessor or colleague without running the risk of confusing materials, wrongly identifying documentation, or not finding it at all.

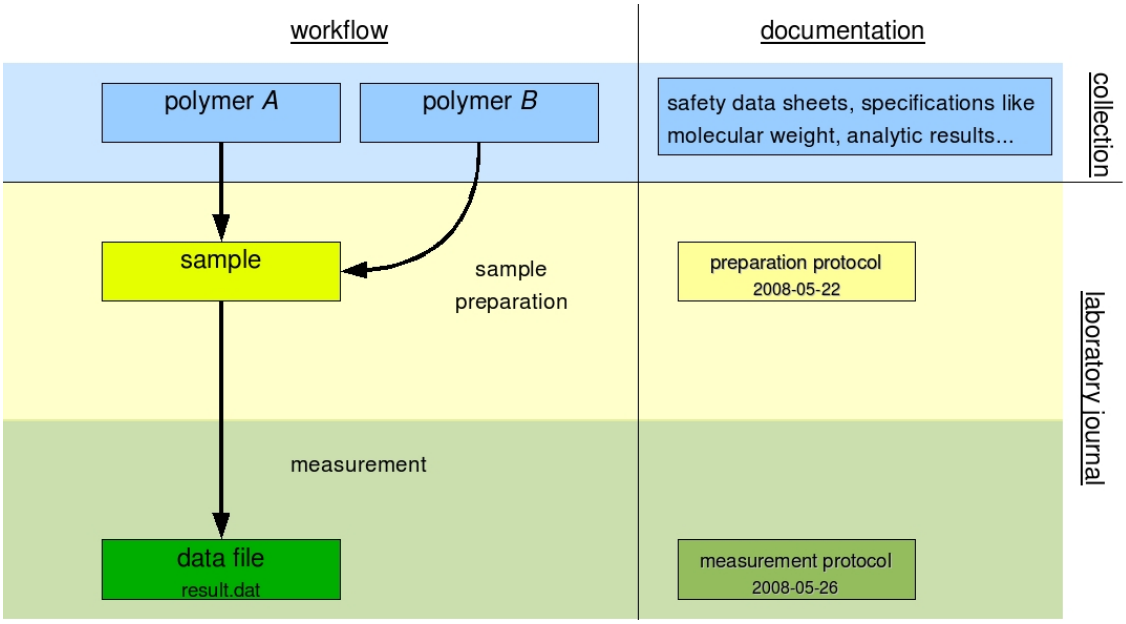

Figure 1. From a substance to measurement data: an example of a typical workflow in materials science. A polymer blend (sample) is created from two polymers (polymer *A* and *B*). The sample is characterized by measuring a characteristic physical property. The result of the measurement is documented in a primary data file (result.dat) and a measurement protocol. Information on the polymers such as safety data sheets, specifications like molecular weight, and analytic results is archived in centralized analog or digital files. The relationship among all the involved objects and information bits is completely clear to the scientist in charge, but other persons will hardly be able to reconstruct the context of a specific primary data file.

3 SYSTEMATIC INFORMATION MANAGEMENT

To avoid the problems described above, two basic principles have to be observed: uniqueness and reproducibility. This means it has to be ensured that all identifiers of materials, samples, and data files are absolutely unique and cannot be mixed up and that the history of each sample or data file can be fully backtracked. The optimal solution would, of course, be to achieve both these objectives in one single step with as little interference as possible with the individual method of information management.

The first condition, viz. unique identifiers, is very easy to meet and only requires some care and discipline in managing the data, labeling samples, and documenting the work processes. There are many possibilities for unique identification. The easiest way is certainly to use a serial number (Ebel et al., 2006), but other methods such as Uniform Resource Locators (URL, Berners-Lee et al., 1994), Uniform Resource Names (URN, Moats, 1997), or hashes (enhanced MD5, Zimmermann et al., 2007) are also suitable.

To ensure backtracking, the documentation must be absolutely consistent with the identification scheme, i.e. the identifier of a data file must be uniquely linked to a measurement protocol. For this purpose, the above techniques are not very suitable to ensure uniqueness. For backtracking, one would, e.g., have to enter the (unique) identifier of the generated data file into a measurement protocol subsequently. Only by doing this would it be possible to link the two items reliably. This involves a considerable amount of subjective extra work, so that this method is not acceptable for the scientific routine. Moreover, scientists are by no means familiar with the use of hashes, URNs, etc., so that this method is simply too complicated because of the care and discipline it requires.

Table 1. Examples of typical timestamps. In most cases appending the timestamp of the daily representation to a keyword will lead to reliable mapping, e.g., of data files to protocols in laboratory journals. Otherwise, the accuracy can be increased or enhanced by additional information like catchwords or the initials of the involved scientist.

timestamp	accuracy	description
2006	annual	The calendar year is suitable if one or not more than a handful of information bits are expected per year. In the latter case the annual timestamp has to be extended by a catchword (German Research Foundation, 2008, p. 28).
2006-02-17	daily	Date representation according to ISO standard 8601 (2004). The year from the Gregorian calendar followed by the month, which is enumerated from 01-12, and the day, which is enumerated from 01-31.
2006-02-17T22:13	every minute	Combined date and time representation according to ISO standard 8601 (2004) with an accuracy of minutes. Date and time are separated by the capital letter T.

There is a method, however, which ensures uniqueness and does not require any additional efforts in the documentation because it is already used every day: the use of timestamps. In the laboratory journal or measurement protocol, the respective date is usually recorded anyway.

Axiom

> The principles of good scientific practice are improved to systematic information management if all identifiers of information bits contain a timestamp of adequate accuracy.

Thus the most reasonable and simple way to ensure uniqueness is to use an identifier composed of a keyword (which everybody can use at his or her discretion) and a timestamp., e.g., keyword polA for polymer *A* (cf. Figure 2) becomes polA-2006-07-03 by adding the delivery date and can now be clearly distinguished from the name of a second, possibly identical material with slightly deviating physical properties from a different supplier. Any risk of confusion is thus eliminated. In the same way, a sample, whose manufacturing date is added to its keyword sam becomes sam-2008-05-22, which is a unique identifier and cannot be mistaken for a sample produced several days later. The same applies to a measurement file, which in this example is named result-2008-05-27.dat and explicitly references the date of measurement. Note that the creation date recorded in electronic file systems is not fixed, but is likely to change, e.g. when the file is copied to a different location.

The use of timestamps has a second advantage: if one uses timestamps in the identifier that are consistent with the documentation (i.e. the manufacturing date in the sample identifier or the measuring date in the data identifier), it is no problem to backtrack the workflow and to clearly assign every piece of documentation, as these timestamps are used in the measurement protocols and laboratory journals anyway and need not be recorded as an extra step. Figure 2 gives an example of such backtracking: if one has a measurement file named result-2008-05-27.dat and does not know exactly what it is, one can quickly access the relevant measurement protocol, in which the date of the measurement is recorded, via the timestamp 2008-05-27 indicated in the identifier. The measurement protocol references the measured sample by means of its unique identifier sam-2008-05-22. If details must be known about the production of the sample in order to be able to interpret the measurement data, the timestamp of the sample name is the link to the corresponding preparation protocol in the laboratory journal, which also provides access quickly and without any problems via the production date. Moreover, the laboratory journal contains details about the chemicals used, and via their unique identifier, one can access the respective documentation without any doubt if the location of the reference library is known. Now the history of the measurement file is completely clear and the data can be used again. If the daily timestamp used in Figure 2 is no longer appropriate because uniqueness is lost due to several similar measurements per day, timestamps with a higher accuracy will be needed (Table 1).

The search procedure is simplest if the creator of the knowledge chain is known. Due to the timestamp approach it would even be possible to scan a couple of relevant laboratory notebooks from different scientists to retrieve the desired information and build the necessary chain of research knowledge because the timestamps nomenclature creates a common frame of reference even for other persons. Thus it is easier to browse the laboratory journals and other documentation of colleagues.

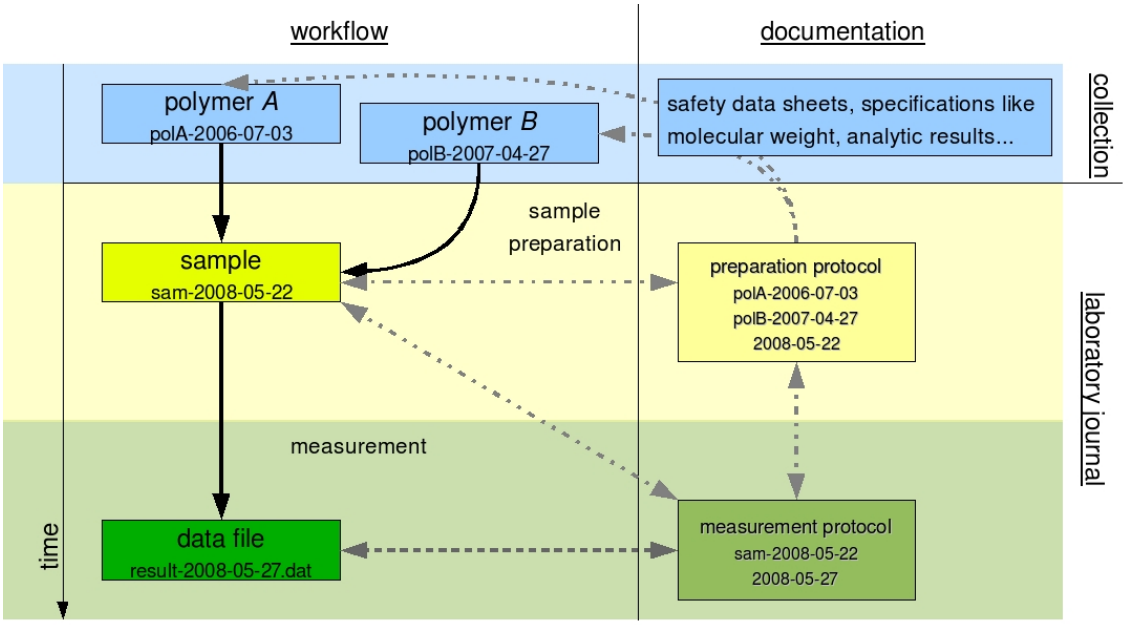

Figure 2. Introducing uniqueness and reproducibility into the traditional information management of natural sciences by means of timestamps. The workflow shown here is an extended version of the workflow presented in Figure 1. Note that each information bit (starting material, sample, or data file) has been given a unique identifier, which is composed of a keyword and a timestamp. This links the information bits uniquely to the documentation of the notebook because the protocols of the laboratory notebook are also dated. Furthermore, the protocols are uniquely linked to the relevant protocols by citing the respective identifiers. For example, the measurement protocol of May 27, 2008, cites the sample sam-2008-05-22 on which the measurement is performed. Similarly, the preparation protocol of May 22 cites the identifiers polA-2006-07-03 and polB-2007-04-27, thus uniquely relating the sample to specific deliveries of the starting materials, which might become important because some material properties may differ to supplier to supplier.

1 DISCUSSION

Although the described naming scheme is suitable for relating measuring files to their experimental history, one sometimes has to search in the opposite direction as well, e.g., to find all information bits derived from a certain sample. On the basis of the discussed axiom, this means that all measurement protocols have to be scanned manually for references to the respective sample to identify the corresponding data files. This is feasible in principle, but certainly takes its time and can be pretty annoying.

If one restricts oneself to using only one unique keyword in all identifiers of one chain of knowledge, one can easily avoid such discomforts. This gives the user much greater search flexibility, which is illustrated in Figure 3 as the extended version of Figure 2. In this example, two different polymers *A* and *B* are processed into two different samples *C* and *D*, and both samples are characterized by individual measurements *E* and *F* (black arrows in

Figure 3). Here, we have two chains of knowledge emanating from samples *C* and *D*. Therefore, in each chain, unique keywords (samC and samD) are used to compose the identifiers of all information bits.

The important point is that this use of a single unique keyword for all identifiers in a chain of knowledge enables a scientist to search an electronic file system for all measurement files related to a specific sample in a really fast way, e.g., by applying the search string samC* (visualized by the dashed arrow between sample *C* and data file *E* in Figure 3).

This is only one example of how the use of identifiers containing a standardized keyword can accelerate and simplify information management. Readers should therefore feel free to adjust the discussed methods to their own individual workflow and requirements. For example, for work groups with a large number of members, it might be suitable to use the initials of the scientist's name as part of the standardized keyword. In this way, the information bits related to a specific scientist can be easily identified.

There is no doubt that managing large quantities of scientific data needs flexible and efficient IT infrastructures (Talbott et al., 2005; Butler, 2005), such as DataFinder (Schlauch & Schreiber, 2007). However, for small work groups in physics, chemistry, biology, or engineering departments or even for students, such infrastructures require too much overhead management. As a consequence, most work groups rely on the traditional method of scientific information management, which has not offered any simple facility up to now to manage analog and electronic information bits and their documentation in parallel. Thus, the presented approach of a simple, but systematic naming scheme for all kinds of information bits solves many problems of daily scientific work because the compilation of unique identifiers from a keyword plus a timestamp enables the scientist to carry on with an individual referencing system. The timestamp introduces a common and intuitive frame of reference, which enables other persons to reconstruct the information flow. Therefore the elementary principles of uniqueness and reproducibility improve the traditional information management of natural sciences to systematic information management.

In fact, the concept of unique identifiers for electronic files as such is not new but is the foundation of the World Wide Web (WWW), which basically depends on Uniform Resource Locators (URL, Berners-Lee et al., 1994) or their replacement, the Uniform Resource Identifier (URI, Berners-Lee et al., 2005). Due to their length and taking into account that URIs are likely to change, important electronic documents are often referenced with Uniform Resource Names (URN, Moats, 1997) or Digital Object Identifiers (DOI, Klump et al., 2006). Of course, the presented naming scheme of the keyword-timestamp type also fits into the WWW because the identifier of a data file, e.g., the filename samC-2008-05-27.dat, can be extended to an URI by simply adding as a prefix the exact location on a file server and a relevant communication protocol such as the hypertext transfer protocol (http, Fielding et al., 1997). On the other hand, modern operating systems enable a user to easily find the location of a primary data file on the basis of a unique file name. For referencing a data file in the local context, it is therefore not necessary to know the exact location. The unique identifier is sufficient, whereby uniqueness can always be guaranteed by choosing an appropriate accuracy for the timestamp (Table 1) or defining that the identifier has to include additional information such as the initials of the scientist's name.

This is the point where the use of unique keywords in research knowledge chains emerges because being able to retrieve all information bits, e.g., of a certain scientist by searching a file system for data file names containing his initials, reveals his personal knowledge chain. Of course, the knowledge chain of a scientist is located on a higher level than the ones related to individual samples (Figure 3), but the example shows that knowledge chains can be arranged hierarchically. For example, they can be used to compile index catalogues of primary data. With their help, figures in monographs can be directly referenced with the corresponding primary data (Liehr, 2003, Appendix A).

Within the data-information-knowledge-wisdom hierarchy (Ackoff, 1989), a research chain of knowledge is located between the information and knowledge levels. It is more than an information bit because it includes the interrelation between several information bits and their appropriate documentation. It is only the source of explicit scientific knowledge, which usually culminates in a scientific publication. It has to be mentioned also that negative research results are often considered not worthwhile for publication by the scientists themselves (Knight, 2003).

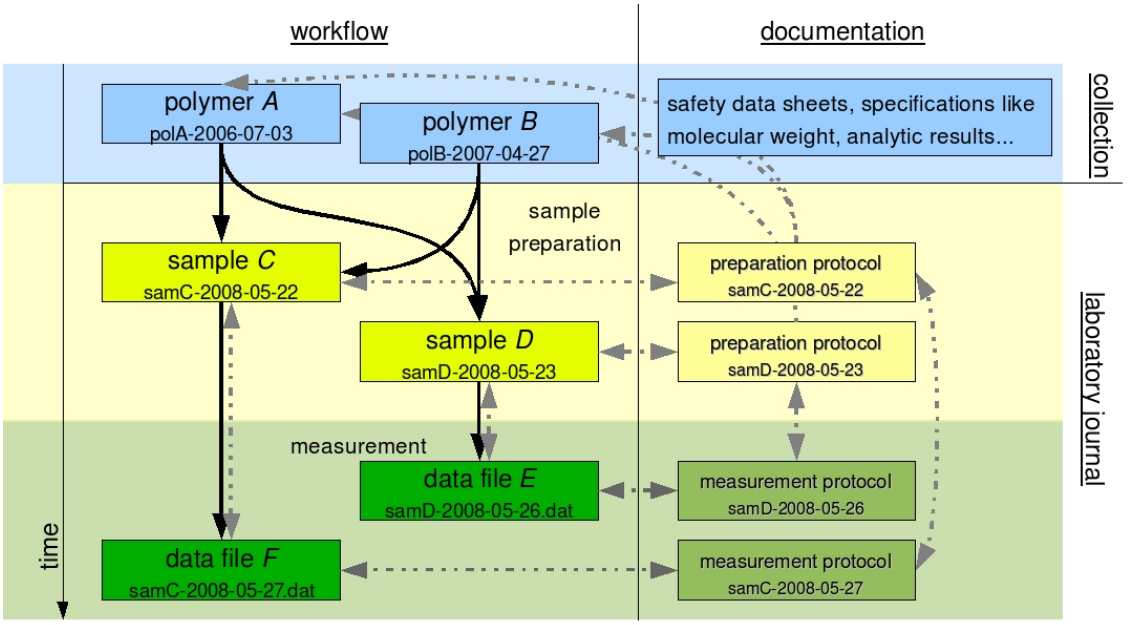

Figure 3. Systematic information management by means of knowledge chains. In this example, two different samples *C* and *D* are prepared from polymers *A* and *B* (black arrows). Each sample is characterized by a measurement (*E* and *F*). The results of the measurements are documented in electronic data files and measurement protocols. Each sample is the origin of a knowledge chain identified by a unique keyword samC and samD, respectively. All interrelated information bits share the same keyword. In combination with electronic file systems, in which the measurement data are stored, this means that all information bits belonging to a certain knowledge chain can be retrieved by searching for the respective keyword of the identifier.

2 CONCLUSION

The purpose of this paper is to present a method that makes it possible in terms of sustainable scientific work to uniquely identify data over prolonged periods of time and to allow the backtracking of their origin. This objective is achieved by consistently using timestamps in all names within the workflow, so that all identifiers are unique and linked to each other. The big advantage of the use of timestamps is that they can be easily added to the current workflow: In the documentation of measurement protocols and digital or analogous laboratory journals, timestamps are used anyway. The naming system is not significantly changed by adding them to the individual names. The individual naming system of a staff member with its respective specific abbreviations thus remains largely unimpaired. Consequently, it is not necessary to develop an additional IT infrastructure, but the method presented here can be implemented immediately and effectively.

The method requires careful and disciplined documentation, but this is the basis of all scientific work anyway. It is primarily suitable for small groups but can also be applied to larger work groups. The basic principles to be observed remain the same: thorough documentation by means of analog or electronic files, unique identification of all information bits, and reliable reproducibility of the scientific knowledge chain.

In view of the simplicity of the discussed information management system, the concept should be included into the curriculum of students or even pupils conducting their first experiments in natural science courses.

3 ACKNOWLEDGEMENTS

The authors would like to thank M. C. Röttger for the fruitful discussions on the topic.

4 REFERENCES

Ackoff, R. L. (1989) From data to wisdom. *Journal of Applied Systems Analysis 16*, 3-9.

Belenkaia, L., Bohnert, M., & Liehr, A. W. (2006) Electronic laboratory notebook assisting reflectance spectrometry in legal medicine. arXiv:cs.DB/0612123.

Berners-Lee, T., Fielding, R., & Masinter, L. (2005) Uniform resource identfier (URI): General syntax. Request for Comment RFC 3986, Network Working Group. Retrieved from the WWW December 12, 2008: http://tools.ietf.org/html/rfc3986

Berners-Lee, T., Masinter, L., & McCahill, M. (1994) Uniform resource locators (URL). Request for Comment RFC 1738, Network Working Group. Retrieved from the WWW December 12, 2008: http://tools.ietf.org/html/rfc1738

Buneman, P., Khanna, S., Tajima, K., & Tan, W.-C. (2004) Archiving scientific data. *ACM Transactions on Database Systems 27 (1)*, 2-42.

Butler, D. (2005) Electronic notebooks: A new leaf. *Nature 436*, 20-21.

Denning, P. J. (2007) Computing is a natural science. *Communications Of The ACM 50 (7*), 13-18.

Ebel, H. F., Bliefert, C., & Greulich, W. (2006) *Schreiben und Publizieren in den Naturwissenschaften, 6th Edition*. Wiley-VCH, Weinheim.

European Commission (2005) The European Charta for Researchers - The Code of Conduct for the Recruitment of Researchers. Tech. Rep. EUR 2160, European Communities. Retrieved from the WWW December 12, 2008:http://ec.europa.eu/eracareers/pdf/am509774CEE_EN_E4.pdf

Fielding, R., Irvine, U., Gettys, J., Mogul, J., Frystyk, H. & Berners-Lee, T. (1997) Hypertext transfer protocol - http/1.1. Request for Comment RFC 2068, Network Working Group. Retrieved from the WWW December 12, 2008: http://tools.ietf.org/html/rfc3986

German Research Foundation (2008) Research grants. DFG-Form 1.02e - 5/08 - II 3, DFG. Retrieved from the WWW December 12, 2008: http://www.dfg.de/forschungsfoerderung/formulare/download/1_02e.pdf

Hartnett, E. & Rew, R. K. (2008) Experience with an enhanced netCDF data model and interface for scientific data access. In: 24th Conference on IIPS. American Metereological Society, New Orleans, p. 7C.2. Retrieved from the WWW December 12, 2008 http://www.unidata.ucar.edu/software/netcdf/papers/AMS_2008.pdf

Herb, U. 2008 Pimp my Repository - *Dritte Open Repositories Conference*. c't (10), 36-37.

Hodgson, J. 2001 The headache of knowledge management. *Nature Biotechnology 19*, 44-46.

ISO 8601 (2004) Data elements and interchange formats - information interchange - representation of dates and times. ISO 8601(E), International Organization for Standardization.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., & Wächter, J. (2006) Data publication in the open access initiative. *Data Science Journal 5*, 79-83.

Knight, J. (2003) Negative results: Null and void. *Nature 422*, 554-555.

Liehr, A. W. (2003) *Dissipative Solitonen in Reaktions-Diffusions-Systemen*. Dissertation, Institut für Angewandte Physik, Westfälische Wilhelms-Universität Münster.

McDowall, R. D. (1999) The role of laboratory information management systems (LIMS) in analytical method validation. *Analytica Chimica Acta 391*, 149-158.

Mlynek, J. (2007) Eröffnungsrede zur Jahrestagung 2007 der Helmholtz-Gemeinschaft: Helle Köpfe fü;r die Forschung. In: Jahrestagung 2007 der Helmholtz-Gemeinschaft. Berlin. Retrieved from the WWW December 12, 2008: http://www.helmholtz.de/aktuelles/reden/artikel/4/4192/

Moats, R. (1997) URN syntax. Request for Comment RFC 2141, Network Working Group. Retrieved from the WWW December 12, 2008: http://tools.ietf.org/html/rfc2141

NIST (2005) NIST chemistry webbook. *NIST Standard Reference Database 69*, National Institute of Standards and Technology. Retrieved from the WWW December 12, 2008: http://webbook.nist.gov/chemistry/

Papiani, M., Wason, J. L., Dunlop, A. N., & Nicole, D. A. (1999) A distributed scientific data archive using the web, xml and sql/med. *SIGMOD RECORD 28 (3*), 56-62. Retrieved from the WWW December 12, 2008: http://eprints.ecs.soton.ac.uk/893/

Polonsky, A., Six, A., Kotelnikov, M., Polonsky, V., Polly, R., & Brey, P. (2006) Semantic laboratory notebook: Managing biomedical research notes and experimental data. In: Leger, A., Kulas, A., Nixon, L., Meersman, R. (Eds.), How Business application challenges Semantic Web Research. *Vol. 194 of CEUR Workshop Proceedings.* Retrieved from the WWW December 12, 2008: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-194/paper5.pdf

Riede, M. K., Sylvester-Hvid, K. O., Glatthaar, M., Keegan, N., Ziegler, T., Zimmermann, B., Niggemann, M., Liehr, A. W., Willeke, G., & Gombert, A. (2008) High throughput testing platform for organic solar cells. *Progress in Photovoltaics: Research and Applications. 16(7)*, 561-576.

Schlauch, T. & Schreiber, A. (2007) Datafinder a scientific data management solution. In: PV 2007. München. Retrieved from the WWW December 12, 2008: http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=F018B97EB1DDE8E920D539ACCFC30253?doi=10.1.1.93.8957

Signaling Gateway (2008) *Molecule pages: A comprehensive signaling database*. Retrieved from the WWW December 12, 2008: http://www.signaling-gateway.org/molecule/

Talbott, T., Peterson, M., Schwidder, J., & Myers, J. D. (2005) Adapting the electronic laboratory notebook for the semantic era. In: McQuay, W., Smari, W. W., Kim, S.-Y. (Eds.), *Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems.* IEEE Computer Society.

Tuchbreitner, A., Marquardt, J., Kappler, B., Honerkamp, J., Kristen, M. O., & Mülhaupt, R. (2003) High-output polymer screening: Exploiting combinatorial chemistry and data mining tools in catalysts and polymer development. *Macromol. Rapid. Comm. 24 (1*), 47-62.

Zimmermann, K., Quack, L., & Liehr, A. W. (2007) Pyphant - a Python framework for modelling reusable information processing tasks. *The Python Papers 2 (3)*, 28-43.