Contribution of Statistics and Value of Data for the Creation of Result Matrices from Objects of Knowledge Resources

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

***Abstract*—This article presents and summarises the main research results on computing optimised result matrices from the practical creation of knowledge resources. With this paper we introduce the main implemented long-term multi-disciplinary and multi-lingual knowledge resources' means, fundamentals and application of documentation, structure, universal classification, and statistics and components for computational workflows and result matrix generation. The resources and workflows can benefit from High End Computing (HEC) resources. The paper presents a knowledge processing procedure using long-term knowledge resources and introduces the n-Probe Parallelised Workflow for an exemplary case study and discussion on a practical application. The goal of this research is to extend the applied features used with long-term knowledge resources' objects and context. The extensions are concentrating on structure and content as well as on processing. The focus is the contribution of statistics and the value of data for the creation of complex result matrices. The major outcome within the last years is the impact on long-term resources based on the scientific results regarding the systematics and methodologies for caring for knowledge.**

***Keywords*–Knowledge Resources; Processing and Discovery; n-Probe Parallelised Workflow; Universal Decimal Classification; High End Computing.**

I. Introduction

Within the last decades the value of data has steadily increased and with this the demand for flexible and efficient discovery processes for creating results from requests on data sources. The fundamental research on optimising result matrices and statistics has been published and presented at the INFOCOMP conference in June 2014 in Paris [1]. This article presents the extended research, especially focussing on data aspects and practical workflows.

Comparable to statistical models used for on-line text classification [2] even more sophisticated models can be used with advanced, structured, and classified knowledge. These models can be assisted using statistical approaches for data analysis [3] in complex information systems as well as for measuring the reliability of classifications models [4] from the content side. The demand for long-term sustainability of the resources increases with the complexity of content and context. The organisation and structure of the resources are getting essentially important, the more important the more the data sizes and complexity as well as their intelligent use are required [5].

The article therefore introduces and discusses the background, including the systematics and methodologies required for an advanced long-term documentation, which can be deployed in most flexible ways – supported by a comprehensive knowledge definition. The general requirements have to consider the condition that it is not sufficient to support only an isolated or special methodology. The knowledge requires special qualities in order to be usable as well as the quantities of knowledge counts. A suitable general conceptual handling and a universal knowledge definition is required in this environment for supporting advanced workflows in benefit for higher qualities of resulting context and matrices. One the side of methodologies and statistics, some major instruments have been developed and successfully integrated. The combination of instruments and resources allows to flexibly compute optimised result matrices for discovery processes in information systems, expert and decision making system components, search engine algorithms, and last but not least supports the further development of the long-term knowledge resources. The presented results are the outcome of the developments and case studies conducted over the last years.

This paper is organised as follows. Section II discusses the motivation, Section III introduces the available knowledge resources regarding processing, workflows, value of data, and their needs for classification and computing. Section IV presents the details of methodologies and components used as it illustrates the details of the implemented resources' features and procedures, structure and classification, statistics. Section V illustrates the resulting, implemented workflow algorithms, and an example for a parallelised workflow, data-centric parallelisation results, and weighted results from statistics and value of data contributions. Section VI discusses the prominent statistics available and tested with the resources and Section VII shows the implementation results for the matrices on a sample case. Sections VIII and IX evaluate the main results and conclude the presented implementation, also discussing the future work.

II. Motivation

Knowledge resources are the basic components in complex integrated systems. Their target is mostly to create a long-term multi-disciplinary knowledge base for various purposes. Request and selection processes result in requirements for computing result matrices from the available information and data. Optimisation in the context of result matrices means "improved for a certain purpose". Here, the certain purpose is given by the target and intention of the application scenario, e.g., requests on search results or associations. Therefore, improving the result matrices is a very multi-fold process and "optimising result matrices" primarily refers to the content and context but in second order also to the workflows and algorithms. The major means presented here contributing to the optimisation are classification and statistics, based on the knowledge resources. The employed knowledge resources can provide any knowledge documentation and additional information on objects and knowledge references, e.g., from natural sciences and decision making. Any data used in case studies is embedded into millions of multi-disciplinary objects, including dynamical and spatial information and data files.

It is necessary to develop logical structures in order to govern the existing unstructured and structured big data today and in future, especially in volume, variability, and velocity and to keep the information addressable on long-term. Preparing and structuring big data is the essential process, which has to preceed creating and implementing algorithms. The systematic, methodological, and "clean" big data knowledge preparation and structuring must generally be named as largest achievement in this context and can be considered by far the most significant overall contribution [5]. The creation and optimisation of respective algorithms is of secondary importance, the more the data must be considered for long-term knowledge creation as, e.g., the benefits of most of those implementations depend on a certain generation of computing and storage architectures, which change all few 4–6 years.

III. Knowledge and resources

With the creation of result matrices we have to introduce a common understanding of knowledge and its processing and the value associated with its application.

A. Knowledge definition and understanding

The World Social Science Report 2013 [6] defines knowledge as "The way society and individuals apply meaning to experience …". Accordingly, the report proposes that "New media and new forms of public participation and greater access to information, are crucial" for open knowledge systems.

In general, we can have an understanding, where knowledge is: Knowledge is created from a subjective combination of different attainments as there are intuition, experience, information, education, decision, power of persuasion and so on, which are selected, compared and balanced against each other, which are transformed and interpreted.

The consequences are: Authentic knowledge therefore does not exist, it always has to be enlived again. Knowledge must not be confused with information or data, which can be stored. Knowledge cannot be stored nor can it simply exist, neither in the Internet, nor in computers, databases, programs or books. Therefore, the demands for knowledge resources in support of the knowledge creation process are complex and multi-fold.

There is no universal "definition" of the term "knowledge", but UDC provides a good overview of the possible width, depth, and facets. For this research the classification references of UDC:0 (Science and knowledge) define the view on universal knowledge [7], which reflects the conceptual dimension and is intended to be used with the full bandwidth of knowledge and knowledge resources.

B. Processing and workflows

Workflows based on the knowledge resources' objects and facilities have been created for different applications. The knowledge resources can make sustainable and vital use of Object Carousels [8] in order to create knowledge object references and modularise the required algorithms [9]. This provides a universal means for improving coverage, e.g., dark data, and quality within the workflow. Secondary resources being available for data, information, and knowledge integration, besides Integrated Information and Computing System (IICS) applications, allow for workflows and intelligent components on High End Computing (HEC) and High Performance Computing (HPC) resources [10], [11]. This paper presents the up-to-date experiences with selected components for structures and workflows.

C. Value of data

The value of data is a central driving force for creating sustainable knowledge resources, the more as data is increasingly important for long period of times. Long-term in cases of sustainable high-value data means many decades of availability and usability. Therefore, usability, security, and archiving are most important aspects of the value of data sets. Value is not the price a data set can be sold as there are many individual factors.

The long-term studies, as the "Cost of Data Breach" study at the Ponemon Institute [12] summarise that the costs related to data loss are high and as predicted [13] do increase [14] every year [15], [16] (sponsored by Symantec), [17] (sponsored by IBM). Straight approaches for calculating individual risks and data loss, as with the Symantec Data Breach Calculator [18] illustrate the effects. Besides science and industry, assessing knowledge loss risks resulting from departing personnel and other factors of loss [19], [20] can be summarised by the risk of knowledge loss, the probability for loss of employees, the consequences of human knowledge loss, and the quality of knowledge resources.

The high quality and value of the knowledge resources used for supporting discovery processes are results of the multi- and trans-disciplinary long-term creation and documentation processes, the structuring of the data, the context of knowledge objects, and the availableness of an universal classification.

D. Knowledge resources

The knowledge resources implement structure and features and can be integrated most flexibly into information and computing system components. Main elements are so called knowledge objects. The objects can consist of any content and context documentation and can employ a multitude of means for description and referencing of objects, data sets, collections, used with computational workflows. Essential core attributes are a facetted universal classification and various content views and attributes, created manually and automated in interactive and batch operation. Developing workflow implementations for various purposes requires to compute result matrices from the knowledge objects and referred knowledge. The purposes can require individual processing means, complex algorithms, and a base of big data collections. Advanced discovery workflows can easily demand large computational requirements for High End Computing (HEC) resources supporting an efficient implementation.

IV. Methodologies and components employed

The following passages refer to the main components and methodologies and introduce the main aspects for the creation of result matrices.

A. Content, context, and procedures

The data used here is based on the content and context from the knowledge resources, provided by the LX Foundation Scientific Resources [21], [22]. The LX structure and the classification references based on UDC [23], [24] are essential means for the processing workflows and evaluation of the knowledge objects and containers. Both provide strong multi-disciplinary and multi-lingual support. The analysis of different classifications and development of concepts for intermediate classifications from the Knowledge in Motion (KiM) long-term project [25] has contributed to the application of UDC in the context of knowledge resources.

An instructive example for an archaeological and geoscientific use case, deploying knowledge resources, classification, references, and Object Carousels has been recently published [8]. With this research the presentation complements the use case by an important methodology, statistics for intermediate result matrices, usable in any associated workflow. In order to get an overview, the following practical example for a specific workflow as part of an application component shows how result matrices for requests can be computed iteratively.

1) Application component request,
2) Object search (i.e., knowledge objects, classification, references, associations),
3) Creation of intermediate result matrices,
4) Iterative and alternating matrix element creation (i.e., based on intermediate result matrices, object search, referenced content, classification, and statistics),
5) Creation of result matrix,
6) Application component response.

The workflow will mostly be linear if the used algorithms are linear and the data involved is fixed in number and content.

The knowledge objects are under continuous development for more than twenty-five years. The classification information has been added in order to describe the objects with the ongoing research and in order to enable more detailed documentation in a multi-disciplinary and multi-lingual context.

Classification is state-of-the-art with the development of the knowledge resources, which implicitly means that the classification is not created statically or even fixed. It can be used and dynamically modified on the fly, e.g., when required by a discovery workflow description. Representations and references can be handled dynamically with the context of a discovery process. So, the classification can be dynamically modelled with the workflow context. The applied workflows and processing are based on the data and extended features developed for the Gottfried Wilhelm Leibniz resources [26].

Mathematical statistics is a central means for data analysis [27], [28]. It can be of huge benefits when analysing regularities and patterns when used for machine learning with information system components [29]. It is a valuable means deployed in natural sciences and has been integrated in multi-disciplinary humanities-based disciplines, e.g., in archaeology [30]. The span of fields for statistics is not only very broad but statistics itself goes far beyond a simple “tool” status [31].

Methodological means, which have been created in order to be deployed for regular use are workflows improving result quantity and result quality, various filters, universal classifications, statistics applications, manually documented resources’ components, integration interfaces for knowledge resources, comparative methods, combination of several means.

The methodologies with the knowledge resources are based on computational methods, processing, classification and structuring of multi-disciplinary knowledge, systematic documentation, long-term knowledge creation, vitality of data concepts, sustainable resources architecture, and collaboration frameworks.

In the past, many algorithms have been developed and implemented [21], [22] for supporting different targets, e.g., silken criteria, statistics, classification, references and citation evaluation, translation, transliteration, and correction support, regular expression based applications, phonetic analysis support, acronym expansions, data and application assignments, request iteration, centralised and distributed discovery, and automated and manual contributions to the workflow.

B. Structure and classification

The key issues for computing result matrices from knowledge resources are that they require long-term tasks on efficiently structuring and classifying content and context. The classification, which has shown up being most important with complex multi-disciplinary long-term classification with practical simple and advanced applications of knowledge resources is the Universal Decimal Classification (UDC) [32].

According to Wikipedia currently about 150,000 institutions, mostly libraries and institutions handling large amounts of data and information, e.g., the ETH Library (Eidgenössische Technische Hochschule), are using basic UDC classification worldwide [33], e.g., with documentation of their resources, library content, bibliographic purposes on publications and references, for digital and realia objects. Just regarding the library applications UDC is present in more than 144,000 institutions and 130 countries [34]. Further operational areas are author-side content classifications and museum collections.

UDC allows an efficient and effective processing of knowledge data. UDC provides facilities to obtain a universal and systematical view on the classified objects. UDC in combination with statistical methods can be used for analysing knowledge data for many purposes and in a multitude of ways.

With the knowledge resources in this research handling 70,000 classes, for 100,000 objects and several millions of referenced data then simple workflows can be linear but the more complex the algorithms get the workflows will mostly become non-linear. They allow interactive use, dynamical communication, computing, decision support, and pre- and postprocessing, e.g., visualisation.

The classification deployed for documentation [35] is able to document any object with any relation, structure, and level of detail as well as intelligently selected nearby hits and references. Objects include any media, textual documents, illustrations, photos, maps, videos, sound recordings, as well as realia, physical objects, such as museum objects. UDC is a suitable background classification, for example:

The objects use preliminary classifications for multi-disciplinary content. Standardised operations used with UDC are coordination and addition ("+"), consecutive extension ("/"), relation (":"), order-fixing ("::"), subgrouping ("[]"), non-UDC notation ("*"), alphabetic extension ("A-Z"), besides place, time, nationality, language, form, and characteristics.

C. Statistics implementation for the knowledge resources

A vast range of statistics, e.g., mathematical statistics, can be deployed based on the knowledge resources. The application of mathematical statistics benefits from an increased number of probes or elements. Probes can result from measurements, e.g., from applied natural sciences and from available material. In many cases, without further analysis a distribution or result may seem random. If the accumulation of an occurrence may indicate a regularity or a rule then this may correlate with a statistical method. Many cases require that statistical results have to be verified for realness. This can be done checking against experience and understanding and using mathematical means, e.g., computing probabilities based on probes.

Statistics have been used for steering the development of the resources. Classification and keyword statistics support the optimisation of the quality of data within the knowledge resources. Counts of terms, references, homophones, synonyms and many more support the improvement of the discovery workflows. Comparisons of content with different language representations increase the intermediate associated result matrices for a discovery process.

The created knowledge resources' architecture is very flexible and efficient because the components allow a natural integration of multi-disciplinary knowledge. The processes of optimising a result matrix differ from a statistical optimisation by the fact that statistics is only one of the factors within the workflows.

V. IMPLEMENTED KNOWLEDGE RESOURCES' MEANS

The goals for the combination of statistics and classification are, for example:

- Creating and improving result matrices.
- Decision making within workflows.
- Further development of knowledge resources.
- Extrapolation and prediction.

The implementation for the required flexible workflow creation and levels is shown in the following sketch (Figure 1).

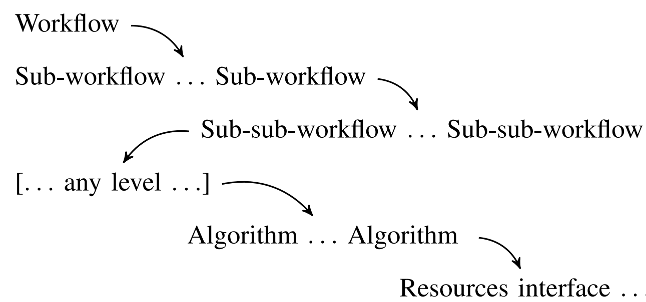

Figure 1. Workflow-algorithm sketch of the implementation (non-hierarchical): Workflow chains, algorithm calls, and resources' interfaces.

The architecture is non-hierarchical. Any workflows can be applied in chains. Each workflow can use sub-workflows, these can use sub-sub-workflows and so on. Each workflow can call or implement algorithms, e.g., for discovery processes, evaluation, and statistics. The workflows and algorithms can use or implement interfaces to the resources. The ellipses indicate that any step can be called or executed in parallel on HEC resources, e.g., in data-parallel or task-parallel processes, in any number of required instances.

An example for this is a "multi-probe parallelised optimisation" workflow, which generates an intermediate result matrix and uses the elements in order to create additional results, all of which are combined for an overall optimised result matrix. The intermediate result matrices are deploying statistical, numerical methods, and various algorithms on base of additional knowledge and information resources.

The knowledge resources allow to implement non-hierarchical and hierarchical architectures. Depending on the workflows these architectures may be created dynamically. Figure 2 shows a workflow-algorithm sketch of a hierarchical implementation based on the resources and emphasizing the methodological aspects.

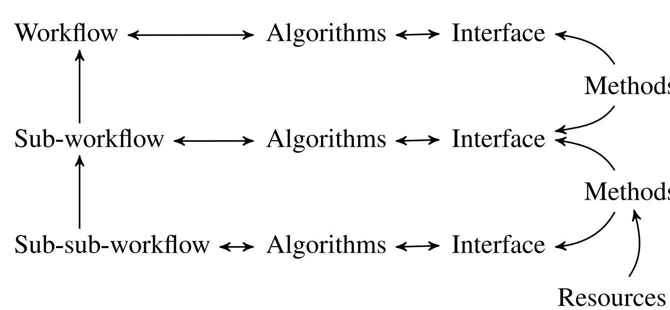

Figure 2. Workflow-algorithm sketch of a hierarchical implementation: Hierarchies of workflows, resources provided by methods.

In this scenario workflows are implemented in a hierarchy of sets workflows, sub-workflows, sub-sub-workflows and so on. Algorithms can be employed by each of these workflows on any level of this hierarchy. The algorithms in turn are connected to the resources through interfaces. The resources can be provided by creating different methods (e.g., static access, dynamic access, batch operation).

A. n-Probe Parallelised Workflow

Computing result matrices can be handled in a multitude of ways. An illustrating example is the n-Probe Parallelised Workflow (nPPW). The workflow is defined by the following steps:

1) A request is started searching for a term called start-element.
2) The search delivers a number of resulting elements, called primary result-elements, being in context with the start-element.
3) The primary result-elements are sorted by a defined attribute (e.g., number of appearance or quality marker).
4) The n most prominent primary result-elements as from the previous step are retained.
5) Secondary requests are started with each of the prominent primary result-elements from the last step.
6) The n most prominent secondary result-elements are gathered for each request, according to the procedure for the primary result-elements.

The workflow is not limited to a single type of elements. Elements can be terms, numbers or other items depending on the use case. For the same reason there is neither a limitation on how to select or weight the elements or which algorithms to use.

The following sketch (Figure 3) demonstrates this at the example of a 5-probe parallelised workflow used for optimisation. In principle, the probes can consist of any type of object, in this example, terms ("T") are used, which are represented by text strings for illustration. c indicates the count for an element in the respective instance while in this example, the absolute count is not in the focus. n indicates the position of the elements.

The "flat search" results in a primary result matrix (Figure 3a) containing terms corresponding with the request for Term 1. In this 5-probe case the resulting primary matrix M0 consists of five elements, Term 1 to Term 5 (dark blue colour).

The "iterative parallelised search instances" in turn get the elements of the result matrix from the flat search as starting seed. In this case, 5-probe means that besides the flat search another four secondary search instances have to be created.

The results of the four secondary requests are secondary result matrices, here, Result Matrix 1 (M1) to Result Matrix 4 (M4) (Figures 3b to 3e). The terms are indexed "Term (m, n)" in short "T (m, n)" with result matrix index m and matrix element index n, starting on the primary matrix at zero.

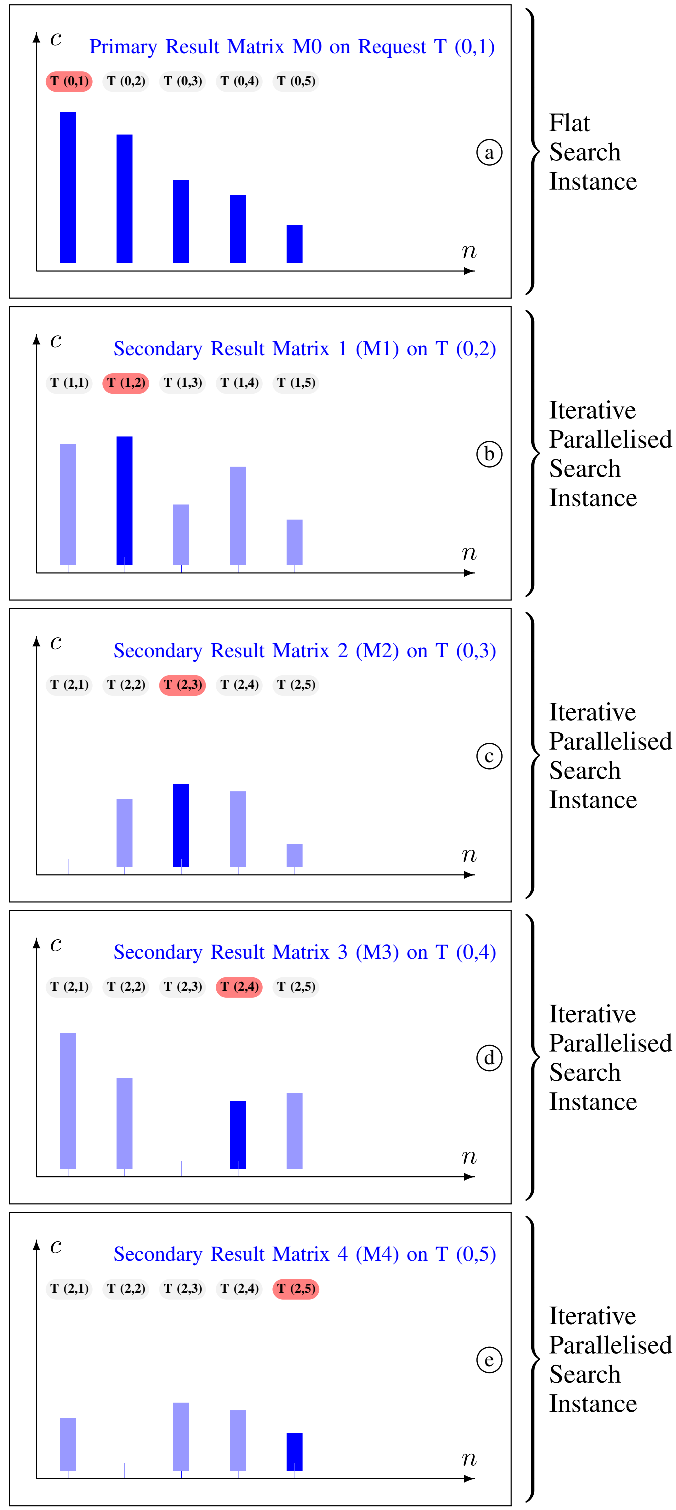

Figure 3. Result matrix creation from a single sub-sub-workflow via intermediate matrices (5-probe parallelised optimisation).

Only those secondary result elements fitting with the original primary elements are considered. The results of the secondary instances are shown in light blue colour. The counts on the various terms differ significantly. Also some secondary search instance can deliver higher counts on a term than the primary search. The larger the primary result matrix is, the higher the number of required consecutive secondary iterative search instances is. In most cases it is a good approach to parallelise the secondary instances, e.g., depending on the available compute resources. The sum of the secondary instances contribute to the overall workflow with an approachingly linear parallelisation curve for an increasing number of instances. As shown, this approach allows statistical support for the iterations, sub-workflows, and discovery algorithms.

The knowledge resources can contribute to the processes and optimisation with increased numbers of objects and also more structured higher quality data included in the processes. In many cases, e.g., for factual knowledge, manually created components provide the highest values with the optimisation. Hybrid "semi-automatically" and automatically created components especially contribute due to their number, dynamical content, and properties.

B. Data-centric parallelisation results

Common workflows can contain an arbitrary number of result matrix operations. In this simple case the matrix contains 5×5 count elements, which may consist of 5 to 21 different terms. As we want to discuss an elementary set of matrix operations every other operations as considered to be pre- and postprocessing in this case:

- Preprocessing workflow,
- Set of result matrix operations,
- Postprocessing workflow.

The calculation depends on the assumption that the resources can provide a sufficient number elements on a specific request via the workflow algorithms.

The following summary (Table I) shows the consequences with n-probe result matrix operations for different numbers n of elements, with $n_{\max} = (n-1)^2 + n$:

TABLE I. N-PROBE: CONSEQUENCES WITH RESULT MATRIX OPERATIONS FOR DIFFERENT NUMBERS n OF ELEMENTS (5, 10, AND 100,000).

Matrix	*Different Elements*	*Parallelisation ⇒ opt. time fact.*
5×5	5–21	5e; 4c ⇒ n:2
10×10	10–91	10e; 9c ⇒ n:2
100,000×100,000	100,000–9,999,900,001	100,000e: 99,999c ⇒ n:2

That means, the algorithm provides a core set of elements and a larger outer race set of elements, which absolutely and relatively increases with increasing matrix sizes.

For a certain implementation allowing soft criteria for the result matrices the relative and absolute numbers and content of the core and outer race set of elements can be adapted in order to create an implementation scalable in terms of data, architecture, operation. In the example presented here (Figure 3) 5 and 16 are particular numbers.

The "core" cores are a reasonable set of cores, which will contribute to the efficiency of the respective result matrix operation. The outer race cores can be handled very flexible. While different distributions of core and outer race sets can still deliver the same results, e.g., for a given set of knowledge resources, they can especially contribute to the workflow scalability and optimisation process. The parallelisation of n elements ("e") with $n-1$ outer race cores ("c") can improve the speedup from an optimisation time factor n to 2, compared with the non-parallel implementation. For a per-instance-cycle of 1 minute a full multi-parallel cycle takes about 2 minutes. Any lower casts of multitude lead to the respective increase of wall times. Under the assumption that the algorithm is not modified for a set of different constellations of compute resources then the process scales about linear. In general, options for providing computing resources are a fixed number of many cores or a situative number of cores. The workflows in this case can be adapted and react to certain compute and storage architectures, considering the situative "Core number of Cores" and the "Cloud Cores" (2C:2C) for the core and outer race sets. This is even more significant as most workflows can integrate dynamical and intelligent components.

C. Weighted results: Statistics and value

Figure 4 illustrates the weighting of the result elements with the above 5-probe parallelised workflow (Figure 3).

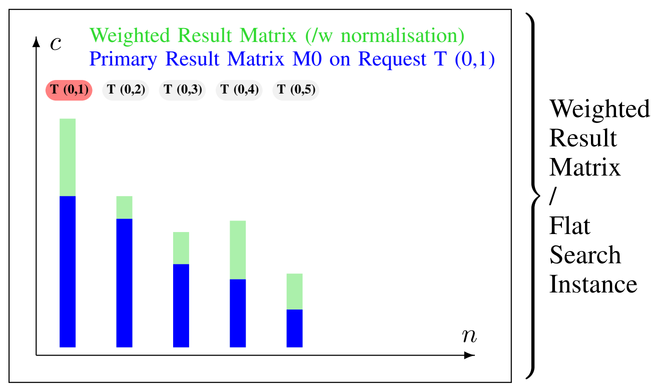

Figure 4. Weighted result matrix (green, without normalisation) creation via intermediate matrices compared to flat search instance (blue), via 5-probe parallelised optimisation.

The weighted result matrix without (/w) normalisation is resulting from the application of the 5-probe parallelised optimisation on the result matrices of the search instances. In this constellation the process is a single sub-sub-workflow, for which we consider the result matrices as intermediate matrices. In contrast to the flat search instance (dark blue colour) the weighted result matrix (green colour) shows different counts. These may consequently result in different priorities and sort orders, as in case of the weighted result for T (0,4) in relation to T (0,3). The weighted priorities and sort orders represent the content and context of the deployed resources, e.g., the asymmetries and references.

The attributes of the content and context can require appropriate algorithms depending of the purposes and workflows for the optimisation. Examples are mean values on counts and fitting to distribution curves with data sets.

D. Statistics and value based on resources

As implementations of statistics are based on counting and numbers the statistics sub-workflows can deploy everything, e.g., any feature or attributes, which can be counted. Sources and means of statistics and computation are:

- Dynamical statistics on the internal and external content and context (e.g., overall statistics, keyword-, categories-, classification-, and media-statistics).
- Mathematics and formula on statistics from the content.
- Elements' statistics (structuring, content, references).
- Statistics based on UDC classification.
- UDC-based statistics computed from comparisons and associations of UDC groups and descriptions.
- Statistics based on any combination of classification, keywords, content, references, context, and computation.

Workflows based on the statistics can be type "semi-manually" or "automated". Besides the major processing and optimisation goals descriptive statistics can be done with each workflow or sub-workflow. Any change of the means supported within a workflow can contribute to the optimisation of the result matrix. Suitable and appropriate means have to be determined for best supporting the goals of the respective step in the workflow. The implementation considers measuring the optimisation by quantity and quality of attributes and features, on intelligence-based and learning processes. With either use there is no general quality measure. Possible quality measures depend on purpose, view, and deployed means. In addition, the decision on these measures can be well supported by statistics, e.g., comparing result matrices from different workflows on the same request. Learning systems components can be used for capturing the success of different measures. The knowledge resources can contain equations and formulary of any grade of complexity. Due to the very high complexity level of the multi-disciplinary components it is necessary to use the basic instances for a comparison in this context of matrix statistics.

The following passages show basic excerpts of statistics objects (LaTeX representation) being part of the implemented knowledge resources. These statistics methods/equations are selected and shown mainly for two reasons: The selected methods are taken from the knowledge objects contained in the resources. These methods are used for result matrix calculations and compared with the evaluation in this research.

VI. Statistics: Fundamentals and Application

Statistics on itself can rarely give an overall decisive answer on a question. Statistic means merely can be used as tools for supporting valuations and decisions. Statistics, probability, and distributions are valuable auxiliaries within workflows and integrated application components, e.g., on numbers of objects, spatial or georeferences, phonetic variations, and series of measurement values. Probability and statistics measures are used with integrated applications, e.g., with search requests, with seismic components (e.g., Median and Mean Stacks), which can also be implemented on base of the resources.

A. Basic algorithms applied with knowledge resources

The mean value, arithmetic mean or average M for n values is given by

$$M = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu \tag{1}$$

Calculating the mean value is described by a linear operation. The median value or central value is the middle value in a size-depending sort order of a number of values. For making a statement on the extent of a group of values, the variance ("scattering") can be calculated, with the mean deviation m and the squared mean deviation m^2.

$$m^2 = \frac{1}{n}\sum_{\nu=1}^{n}(x_\nu - M)^2 = \overline{(x-M)^2} \tag{2}$$

For any value this holds $m^2(A) = m^2 + (M-A)^2$. When applying statistics, especially when calculating the propagated error, the following definition of the variance is used:

$$m^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(x_\nu - M)^2 \tag{3}$$

The mean deviation $\zeta(A)$ is defined as:

$$\zeta(A) = \overline{|x-A|} \text{ for which holds } \zeta(A) = \text{min. for } A = Z \tag{4}$$

The probable deviation or probable error ρ with the probable limits Q_1 and Q_3 is defined as:

$$\rho = \frac{Q_3 - Q_1}{2} \tag{5}$$

The relative frequency h_i is defined as:

$$h_i = \frac{n_i}{n}, \text{ then it holds } \sum_{i=1}^{k} h_i = 1 \tag{6}$$

where n_i is the class frequency, which means the number of elements in a class of which the middle element is x_i.

B. Distributions deployed with knowledge resources

A continuous summation results in the cumulative frequency distribution

$$H_i = \sum_{j=1}^{i} h_j \tag{7}$$

which gives the relative number, for which holds $x \leq x_i \cdot H_i$ is a function discretly increasing from 0 to 1. The presentation results in a summation line. With steady variables, for which at an interval width of Δx the quotient $h_i/\Delta x$ nears a limit, one can calculate a frequency density $h(x_i)$ and for the summation frequency $H(x)$:

$$h(x_i) = \lim_{\Delta x \to 0} \frac{h_i}{\Delta x} \quad \text{and} \quad \frac{dH(x)}{dx} = h(x) \tag{8}$$

With statistical distributions the Gaussian normal distribution is of basic importance.

$$h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{9}$$

$H(x)$ can not be given "closed". It can be shown that

$$K = \int_{-\infty}^{+\infty} h(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx = 1 \tag{10}$$

The Binominal distribution $w_k(s)$ is defined by

$$w_k(s) = \binom{k}{s} p^s q^{k-s} \tag{11}$$

The sum of the two binominal coefficients is equal to $\binom{k+1}{s}$. This is described by Pascals' Triangle. It holds:

$$M = \sum_{s=0}^{k} w_k(s) \cdot s = kp \quad \text{and} \quad m = \sqrt{kqp} \tag{12}$$

Accordingly, the mean error of the mean value decreases proportional to $1/\sqrt{k}$. This describes the error propagation law.

$$h(X) = \frac{1}{\sqrt{2\pi}m} e^{-\frac{1}{2}\left(\frac{X-M}{m}\right)^2} \quad \text{for} \ -\infty < X < +\infty \tag{13}$$

From this Gaussian curves, binominal distributions, correlation coefficients and advanced measures can be developed.

C. Application of fundamental theorems of probability

The probability p is defined by:

$$p(x_i) = \lim_{n\to\infty} h_i = \lim_{n\to\infty} \frac{n_i}{n} \tag{14}$$

The classical definition of p_{classic} is:

$$p_{\text{classic}} = \frac{\text{number of favoured cases}}{\text{number of possible cases}} \tag{15}$$

The following can be said if independence is supposed. $\vee$ means the logical OR, $\wedge$ the logical AND.
Either-OR: If $E_1, E_2, \ldots, E_m$ are events excluding each other and the respective probabilities are $p_1, p_2, \ldots, p_m$, then the probability for *either* E_1 *OR* E_2 *OR* ... *OR* E_m is:

$$p(E_1 \vee E_2 \vee \ldots \vee E_m) = p_1 + p_2 + \ldots + p_m \tag{16}$$

As-well-as: If $E_1, E_2, \ldots, E_m$ are event pairwise independent from each other then the probability of E_1 *as well as* E_2 *as well as* ... *as well as* E_m is:

$$p(E_1 \wedge E_2 \wedge \ldots \wedge E_m) = p_1 \cdot p_2 \cdot \ldots \cdot p_m \tag{17}$$

VII. Implementation for the Result Matrix Cases

A. Measures for optimisation and purposes

The measures for optimisation are on the one hand object of the services and workflows but on the other hand they can be of concern for the knowledge resources themselves.

Conforming with the goals, measures for optimisation mean fitness for a purpose, e.g., search for a regularity with statistics and result matrices. After a search for regularities any statistical procedure benefits from checking against experiences and associating the procedure and result with a meaning. In many cases, e.g., "relevance" means numbers, uniqueness, proximity for objects, content, and attributes, e.g., terms.

Optimisation can be achieved by various means, e.g., by intelligent selection, by self-learning based optimisation, and by comparisons and statistics. The first measures include manual procedures and essences of results being stored for learning processes. They can also deploy comparisons and statistics, which also mean probability and distributions. This case study is focussed on comparisons and statistics applied with the knowledge resources. The subject of the statistics deals with the collection, description, presentation, and interpretation of data. Especially, the methodology can be based on computing more than the minimal number of comparisons, computing more than the minimal number of distributions, computing result matrices considering the mean of several distributions or extreme distributions. In the case of "relevance", information on weighting may come from sources of different qualities.

The general steps with the knowledge resources, including external sources, can be summarised as: Knowledge resource requests, integrating search engine results (e.g., Google), integrating results more or less randomly, without explicit considerate classification and correlation between content and request, comparing the content of search result matrix elements with the knowledge resources result matrix containing classified elements, statistics on an accumulation of terms, selecting accumulated terms, elimination of less concentrated results, selecting the appropriate number of search results.

B. Sources and Structure: Knowledge resources

The full content, structure, and classification of the knowledge resources have been used. In the context of the case discussed here, the sources, which have been integrated and referenced with the knowledge resources consist of:

- Classical natural sciences data sources.
- Environmental and climatological information.
- Geological and volcanological information.
- Natural and man-made factor/event information.
- Data sets and compilations from natural sciences.
- Archaeological and historical information.
- Archive objects references to realia objects.
- Photo and video objects.
- Dynamical and non-dynamical computation of content.

The sources consist of primary and secondary data and are used for workflows, as far as content or references are accessible and policies, licenses, and data security do not restrict.

C. Classification and statistics in this sample case

Table II shows a small excerpt of resulting main UDC classification references practically used for the statistics with the knowledge resources in the example case presented here.

TABLE II. UNIVERSAL DECIMAL CLASSIFICATION OF STATISTICS FEATURES WITH THE KNOWLEDGE RESOURCES (EXCERPT).

UDC Code	*Description*
UDC:3	Social Sciences
UDC:310	Demography. Sociology. Statistics
UDC:311	Statistics as a science. Statistical theory
UDC:311.1	Fundamentals, bases of statistics
UDC:311.21	Statistical research
UDC:311.3	General organization of statistics. Official statistics
UDC:5	Mathematics. Natural sciences
UDC:519.2	Probability. Mathematical Statistics
UDC:531.19	Statistical mechanics
UDC:570.087.1	Biometry. Statistical study and treatment of biological data
UDC:615.036	Clinical results. Statistics etc.

The small unsorted excerpts of the knowledge resources objects only refer to main UDC-based classes, which for this part of the publication are taken from the Multilingual Universal Decimal Classification Summary (UDCC Publication No. 088) [23] released by the UDC Consortium under the Creative Commons Attribution Share Alike 3.0 license [36] (first release 2009, subsequent update 2012).

As with any object the statistics features can be combined for facets and views for any classification subject. On the other hand statistics objects from the resources can be selected and applied. The listing (Figure 5) shows an excerpt intermediate object result matrix on statistics content.

```
ANOVA         [Statistics, ...]:
              Analysis of Variance.
BIWS          [Whaling]:
              Bureau of International Whaling Statistic.
GSP           [Geophysics]:
              Geophysical Statistics Project.
Median        [Statistics]:
              In the middle line.
              s. also Median-Stack
Median-Stack  [Seismics]:
              Stacking based on the median value of
               adjacent traces.
MSWD          [Mathematics]:
              Mean Square Weighted Deviation.
MSA           [Abbbreviation, GIS]:
              Metropolitan Statistical Area.
MOS           [Abbreviation]:
              Model Output Statistics.
MCDM          [GIS, GDI, Statistics, ...]:
              Multi-Criteria Decision Making.
SHIPS         [Meteorology]:
              Statistical Hurricane Intensity Prediction
               Scheme.
SAND          [Abbreviation]:
              Statistical Analysis of Natural resource
               Data, Norway.
```

Figure 5. Intermediate object result matrix on “statistics” content.

Learning from this: The classifications used for this intermediate matrix are based on contributions from more than one discipline. The elements themselves do not necessarily have to contain a requested term because the classification contributes. Several steps may be necessary in order to improve the matrix, e.g., selecting disciplines, time intervals on the entries, references, and associations. Because different content carries different attributes and features the evaluation can be used in comparative as well as in complementary context.

The implemented knowledge resources means of statistics and computation described above are integrated in the workflows, including classification, dating, and localisation of objects. In addition, probability distributions, linear and non-linear modelling, and other supportive tools are used within the workflow components.

D. Resulting numbers on processing and computing

The processing and computational demands per workflow instance result from the implementation scenarios. The following comparison (Table III) results from a minimal workflow request for a result matrix compared to a workflow request for a result matrix supporting classification views referring to UDC, supporting references and statistics on intermediate results. Both scenarios are based on the same number of elements and entries and can be considered atomic instances in a larger workflow. Views and result matrices can be created manually and automated in interactive and batch operation.

TABLE III. PROCESSING AND COMPUTATIONAL DEMANDS: 2 SCENARIOS, BASED ON 50000 OBJECT ELEMENTS AND 10 RESULT MATRIX ENTRIES.

Scenario Workflow Request for Result Matrix	*Value*
“`geosciences archaeology`” (minimal)	
Number of elements	50,000
Number of result matrix entries (defined)	10
Number of workflow operations	15
Wall time on one core	14 s
“`geosciences archaeology`” (UDC, references, statistics)	
Number of elements	50,000
Number of result matrix entries (defined)	10
Number of workflow operations	6,500
Wall time on one core	6,700 s

As the discussed scenarios are instances this means workflows based on n of these instances will at least require n-times the time for an execution on the same system. It must be remembered that the parallelisation will have a significant effect when workflows are created based on many of these instances when required in parallel. Without modifying the algorithms of the instances, which mostly means simplifying, the positive parallelisation effect for the workflows can be nearly linear. Besides the large requirements per instance with most workflows there are significant beneficial effects from parallelising even within single instances as soon as the number of comparable tasks based on the instances increases. A typical case where parallelisation within a workflow is favourable is the implementation of an application creating result matrices and being used with many parallel instances, e.g., with providing services. The number of 70,000 elementary UDC classes currently results in 3 million basic elements when only considering multi-lingual entries – without any combinations. With most isolated resources only several thousand combinations are used in practice each. The variety and statistics are mostly deployed for decision processing, increasing quantity, and increasing quality. Many of the above cases require to compute more than one data-workflow set to create a decision. A review and an auditing process are mandatory for mission critical applications. The computational requirements can increase drastically with the computation of multiple workflows. Each workflow will consist of one or more processes, which can contain different configurations and parameters. Therefore, creating a base for an improved result matrix starts with creating several intermediate result matrices. With a ten process workflow, e.g., the possible configurations and parameters can easily lead to computing a reasonable set of thousands to millions of intermediate result matrices.

The objects and methods used can be long-term documented as knowledge objects. Nevertheless, there is explicitely no demand for a certain programming language. Even multiple implementations can be done with any object. The workflows and algorithms with the cases discussed here have been implemented as objects in Fortran, Perl, and Shell. Anyhow, the implementation of algorithms is explicitely not part of any core resources. It is the task of anyone having an application to do this and to decide on the appropriate means and methods.

E. Complementary Components

As an example we choose to mention three state-of-the-art components for implementing the "data-base", operating system, and distributed platform. With this it should be possible to build and use containers. For implementation of very simple non-hierarchical but data-set centred scenarios the MongoDB [37] concept may be used. This database model greps the concept of a data-set centred approach and extends the traditional database models. CoreOS [38] can be used for data-warehouse style computing, providing and operating system for massive server deployment. In addition, Docker [39] can be used, an open platform for distributed applications, which shall enable to build, ship and run applications anywhere.

Anyhow, these components are not data-centric themselves. It is also more than questionable if data can be sustainably preserved in close integration with these components, even for mid-term purposes of a few decades only.

The knowledge resources, including their creation and further development, should be kept in a long-term and portable concept, as an implementation based on such above components has shown to be still much too application centric.

VIII. Case results and evaluation

Computing result matrices is an arbitrary complex task, which can depend on various factors. Applying statistics and classification to knowledge resources has successfully provided excellent solutions, which can be used for optimising result matrices in context of natural sciences, e.g., geosciences, archaeology, volcanology or with spatial disciplines, as well as for universal knowledge. The method and application types used for optimisation imply some general characteristics when putting discovery workflows into practice regarding components like terms, media, and other context (Table IV).

TABLE IV. RESULTING PER-INSTANCE-CALLS FOR METHOD AND APPLICATION TYPES ON OPTIMISATION WITH KNOWLEDGE DISCOVERY.

Type	*Terms*	*Media*	*Workflow*	*Algorithm*	*Combination*
Mean	500	20	20	50,000	3,000
Median	10	5	2	5,000	50
Deviation	30	5	5	200	20
Distribution	90	40	15	20	120
Correlation	15	10	5	20	90
Probability	140	15	20	50	150
Phonetics	50	5	10	20	50
Regular expr.	920	100	50	40	1,500
References	720	120	30	5	900
Association	610	60	10	5	420
UDC	530	120	20	5	660
Keywords	820	100	10	5	600
Translations	245	20	5	5	650
Corrections	60	10	5	5	150
External res.	40	30	5	5	40

Statistics methods have shown to be an important means for successfully optimising result matrices. The most widely implemented methods for the creation of result matrices are intermediate result matrices based on regular expressions and intermediate result matrices based on combined regular expressions, classification, and statistics, giving their numbers special weight. Based on these per-instance numbers this results in demanding requirements for complex applications – On numerical data: Millions of calls are done per algorithm and dataset, hundreds in parallel/compact numeric routines. On "terms": Hundred thousands of calls are done per sub-workflow, thousands in parallel/complex routines, are done.

Most resources are used for one application scenario only. Only 5–10 percent overlap between disciplines – due to mostly isolated use. Large benefits result from multi-disciplinary multi-lingual integration. The multi-lingual application adds an additional dimension to the knowledge matrix, which can be used by most discovery processes. As this implemented dimension is of very high quality the matrix space can benefit vastly from content and references.

IX. Conclusion and future work

This paper presented the extended research, focussing on data aspects and practical workflows, based on the fundamental research on optimising result matrices from knowledge discovery workflows. This research has extended the applied features used with long-term knowledge resources' objects and context. Starting with the multi-disciplinary and multi-lingual knowledge resources examples for non-hierarchical and hierarchical workflows have been presented.

First, knowledge resources' objects with their structured content, references, and conceptual knowledge are providing an excellent means for long-term multi-disciplinary and multi-lingual documentation and reuse. This especially includes the flexible universal classification of any objects. The quality of data can be used to contribute to the discovery and optimisation processes, which increases the emphasis on the values of data the more the long-term significance gets into the focus.

Second, the use of statistics and algorithms based on statistics has shown to provide solid tools for creating and improving result matrices. Both, the documentation and resources and the statistics applicable in workflows result in benefits for complex result matrix generation.

The case study introduced the application of n-Probe Parallelised Workflows, which can be used for result matrix generation. The matrix generation and processes have been discussed in detail. Workflows like these have been successfully used for the optimisation of result matrices. They allow to use statistics methods and data value weighting and can contribute to the creation and development of resources. A number of structuring elements and workflow procedures have been successfully implemented for processing objects from knowledge resources, which allow optimising result matrices in very flexible ways. Long-term multi-disciplinary and multi-lingual knowledge resources can provide a solid source of structured content and references for a wealth of result matrices. The long-term results confirm that for the usability the organisation of the content and the data structures are most important and should have the overall focus compared to algorithm adaptation and optimisation. Nevertheless, the computational requirements may be very high but compared against the long-term data creation issues, they should be regarded secondary from the scientific point of view. Employing a classification like UDC has shown to be a universal and most flexible solution with statistics for supporting long-term multi-disciplinary knowledge resources. Computing optimised result matrices from objects of universally classified knowledge resources can be efficiently supported by various statistics and probability measures. With the quality and quantity of matrix elements this can also improve the decision making processes within the workflows.

The research conducted provided that advanced discovery will have to go into depth as well as into broad surface of the context of the multi-disciplinary and multi-lingual information in order to effectively improve the quality for most workflows. Many of these workflow processes can be very well parallelised on HEC resources. A typical case where parallelisation is required is the implementation of an application creating result matrices and used with many parallel instances. This introduces benefits for the applicability of the discovery facing big data resources to be included. The integration of the above strategies and means has proven an excellent method for computing optimised result matrices.

On the computational side, the workflows contribute to the parallelisation of processes and result in higher scalability regarding data resources, architectures, and operation. Therefore, the resources and processing workflows can benefit from a flexible deployment of High End Computing resources. The major outcome on the content side is the impact on long-term resources based on the scientific results regarding the systematics and methodologies for caring for knowledge.

Besides all future application scenarios, the further creation of development of content and context and its documentation is a main goal. Future work will be focussed on the workflow processes and standardisation and best practice for container and resources' objects but also concentrate on the development of flexible structures for objects and the automation of processes.

ACKNOWLEDGEMENTS

We are grateful to all national and international partners in the GEXI cooperations for their support and contributions. We thank the Science and High Performance Supercomputing Centre (SHPSC) for long-term support of collaborative research since 1997, including the GEXI developments and case studies on archaeological and geoscientific information systems. Special thanks go to the scientific colleagues at the Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, especially Dr. Friedrich Hülsmann, for collaboration and prolific discussion within the "Knowledge in Motion" project, for inspiration, and practical case studies. Many thanks go to the scientific colleagues at the Leibniz Universität Hannover, especially to Dipl.-Biol. Birgit Gersbeck-Schierholz and to Dr. Bernhard Bandow, Max Planck Institute for Solar System Research (MPS), Göttingen, for their collaboration and discussion on non-hierarchical and hierarchical workflows, as well as to the scientific colleagues at the Institute for Legal Informatics (IRI), Leibniz Universität Hannover, and to the Westfälische Wilhelms-Universität (WWU), for discussion, support, and sharing experiences on collaborative computing and knowledge resources and for participating in fruitful case studies as well as to my students and participants of the postgraduate European Legal Informatics Study Programme (EULISP) for prolific discussion of scientific, legal, and technical aspects over the last years.

REFERENCES

[1] C.-P. Rückemann, "Computing Optimised Result Matrices for the Processing of Objects from Knowledge Resources," in Proceedings of The Fourth International Conference on Advanced Communications and Computation (INFOCOMP 2014), July 20–24, 2014, Paris, France. XPS Press, 2014, pages 156–162, ISSN: 2308-3484, ISBN-13: 978-1-61208-365-0, URL: http://www.thinkmind.org/index.php?view=article&articleid=infocomp_2014_7_20_60039 [accessed: 2015-02-01].

[2] P. Cerchiello and P. Giudici, "Non parametric statistical models for on-line text classification," Advances in Data Analysis and Classification – Theory, Methods, and Applications in Data Science, vol. 6, no. 4, 2012, pp. 277–288, special issue on "Data analysis and classification in marketing" Baier, D. and Decker, R. (guest eds.) ISSN: 1862-5347 (print), ISSN: 1862-5355 (electronic).

[3] W. Gaul and M. Schader, Eds., Classification As a Tool of Research. North-Holland, Amsterdam, 1986, Proceedings, Annual Meeting of the Classification Society, (Proceedings der Fachtagung der Gesellschaft für Klassifikation), ISBN-13: 978-0444879806, ISBN-10: 0-444-87980-3, Hardcover, XIII, 502 p., May 1, 1986.

[4] J. Templin and L. Bradshaw, "Measuring the Reliability of Diagnostic Classification Model Examinee Estimates," Journal of Classification, vol. 30, no. 2, 2013, pp. 251–275, Heiser, W. J. (ed.), ISSN: 0176-4268 (print), ISSN: 1432-1343 (electronic), URL: http://dx.doi.org/10.1007/s00357-013-9129-4 [accessed: 2015-02-01].

[5] A. Woodie, "Forget the Algorithms and Start Cleaning Your Data," Datanami, 2014, March 26, 2014, URL: http://www.datanami.com/datanami/2014-03-26/forget_the_algorithms_and_start_cleaning_your_data.html [accessed: 2015-02-01].

[6] World Social Science Report 2013, Changing Global Environments, 1st ed. Published jointly by the United Nations Educational, Scientific and Cultural Organization (UNESCO), the International Social Science Council (ISSC), the Organisation for Economic Co-operation and Development (OECD), 2013, DOI: 10.1787/9789264203419-en, OECD ISBN 978-92-64-20340-2 (print), OECD ISBN 978-92-64-20341-9 (PDF), UNESCO ISBN 978-92-3-104254-6 (PDF and print).

[7] C.-P. Rückemann, "From Multi-disciplinary Knowledge Objects to Universal Knowledge Dimensions: Creating Computational Views," International Journal On Advances in Intelligent Systems, vol. 7, no. 3&4, 2014, pp. 385–401, ISSN: 1942-2679, LCCN: 2008212456 (Library of Congress).

[8] C.-P. Rückemann, "Sustainable Knowledge Resources Supporting Scientific Supercomputing for Archaeological and Geoscientific Information Systems," in Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal. XPS Press, 2013, pp. 55–60, ISSN: 2308-3484, ISBN: 978-1-61208-310-0, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2012_3_10_10012 [accessed: 2015-02-01].

[9] C.-P. Rückemann, "High End Computing for Diffraction Amplitudes," in The Third Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of The 11th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 21–27, 2013, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), AIP Conference Proceedings, vol. 1558. AIP Press, 2013, pp. 305–308, ISBN: 978-0-7354-1184-5, ISSN: 0094-243X, DOI: 10.1063/1.4825483.

[10] U. Inden, D. T. Meridou, M.-E. C. Papadopoulou, A.-C. G. Anadiotis, and C.-P. Rückemann, "Complex Landscapes of Risk in Operations Systems Aspects of Processing and Modelling," in Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal. XPS Press, 2013, pp. 99–104, ISSN: 2308-3484, ISBN: 978-1-61208-310-0, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2013_5_10_10114 [accessed: 2015-02-01].

[11] P. Leitão, U. Inden, and C.-P. Rückemann, "Parallelising Multi-agent Systems for High Performance Computing," in Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal. XPS Press, 2013, pp. 1–6, ISSN: 2308-3484, ISBN: 978-1-61208-310-0, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2013_1_10_10055 [accessed: 2015-02-01].

[12] Ponemon Institute, "Data-Security," 2014, Ponemon Institute, URL: http://www.ponemon.org/data-security [accessed: 2015-02-01].

[13] C.-P. Rückemann, "High End Computing Using Advanced Archaeology and Geoscience Objects," International Journal On Advances in Intelligent Systems, vol. 6, no. 3&4, 2013, pp. 235–255, iSSN: 1942-2679, URL: http://www.iariajournals.org/intelligent_systems/intsys_v6_n34_2013_paged.pdf [accessed: 2015-02-01].

[14] Ponemon Institute, "Ponemon Study Shows the Cost of a Data Breach Continues to Increase," 2014, Ponemon Institute, URL: http://www.ponemon.org/news-2/23 [accessed: 2015-02-01].

[15] Ponemon Institute, "Cost of Data Breach 2011," 2011, Ponemon Institute / Symantec, URL: http://www.ponemon.org/library/archives/2012/03 [accessed: 2015-02-01].

[16] Ponemon Institute, "2013 Cost of Data Breach: Global Analysis," 2013, Ponemon Institute / Symantec, URL: http://www.ponemon.org/local/upload/file/2013%20Report%20GLOBAL%20CODB%20FINAL%205-2.pdf [accessed: 2015-02-01].

[17] Ponemon Institute, "2014 Cost of Data Breach: Global Analysis," May 2014, Ponemon Institute / IBM, Ponemon Institute (c) Research Report, Benchmark research sponsored by IBM, Independently conducted by Ponemon Institute LLC, IBM Document Number: SEL03027USEN, URL: http://www.ibm.com/services/costofbreach [accessed: 2015-02-01], URL: http://public.dhe.ibm.com/common/ssi/ecm/en/sel03027usen/SEL03027USEN.PDF [accessed: 2015-02-01].

[18] Symantec, "Symantec Data Breach Calculator," 2014, Symantec, URL: https://databreachcalculator.com/ [accessed: 2015-02-01].

[19] "Wissensverlust vermeiden beim Abgang von Wissensarbeitern," library essentials, LE_Informationsdienst, Juni/Juli 2014, 2014, pp. 9–11, ISSN: 2194-0126, URL: http://www.libess.de [accessed: 2015-02-01].

[20] M. E. Jennex, "A Proposed Method for Assessing Knowledge Loss Risk with Departing Personnel," vol. 44, no. 2, 2014.

[21] "LX-Project," 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX (Information) [accessed: 2015-02-01].

[22] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in Proc. INFOCOMP 2012, Oct. 21–26, 2012, Venice, Italy, 2012, pp. 36–41, ISBN: 978-1-61208-226-4.

[23] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2015-02-01].

[24] "UDC Online," 2014, URL: http://www.udc-hub.com/ [accessed: 2015-02-01].

[25] F. Hülsmann and C.-P. Rückemann, "Value of Data and Long-term Knowledge," KiMrise, Knowledge in Motion, August 12, 2014, 10 Year Anniversary Workgroup Meeting, "Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF)", Hannover, Germany, 2014.

[26] C.-P. Rückemann, "Archaeological and Geoscientific Objects used with Integrated Systems and Scientific Supercomputing Resources," International Journal on Advances in Systems and Measurements, vol. 6, no. 1&2, 2013, pp. 200–213, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/download.php?articleid=sysmea_v6_n12_2013_15 [accessed: 2015-02-01], URL: http://lccn.loc.gov/2008212470 [accessed: 2015-02-01].

[27] Y. Dodge, The Oxford Dictionary of Statistical Terms. Oxford University Press, 2006, ISBN: 0-19-920613-9.

[28] B. S. Everitt, The Cambridge Dictionary of Statistics, 3rd ed. Cambridge University Press, Cambridge, 2006, ISBN: 0-521-69027-7.

[29] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006, ISBN: 0-387-31073-8.

[30] R. D. Drennan, Statistics in Archaeology. Elsevier Inc., 2008, in: Pearsall, Deborah M. (ed.), Encyclopedia of Archaeology, pp. 2093-2100, Elsevier Inc., ISBN: 978-0-12-373962-9.

[31] D. Lindley, "The Philosophy of Statistics," Journal of the Royal Statistical Society, 2000, JSTOR 2681060, Series D 49 (3), pp. 293–337, DOI: 10.1111/1467-9884.00238.

[32] "Universal Decimal Classification Consortium (UDCC)," 2014, URL: http://www.udcc.org [accessed: 2015-02-01].

[33] "Universal Decimal Classification (UDC)," 2015, Wikipedia, URL: http://en.wikipedia.org/wiki/Universal_Decimal_Classification [accessed: 2015-02-01].

[34] A. Slavic, "UDC libraries in the world - 2012 study," universaldecimalclassification.blogspot.de, 2012, Monday, 20 August 2012, URL: http://universaldecimalclassification.blogspot.de/2012/08/udc-libraries-in-world-2012-study.html [accessed: 2015-02-01].

[35] C.-P. Rückemann, "Integrating Information Systems and Scientific Computing," International Journal on Advances in Systems and Measurements, vol. 5, no. 3&4, 2012, pp. 113–127, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/index.php?view=article&articleid=sysmea_v5_n34_2012_3/ [accessed: 2015-02-01].

[36] "Creative Commons Attribution Share Alike 3.0 license," 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ [accessed: 2015-02-01].

[37] "MongoDB," 2015, URL: http://www.mongodb.org/ [accessed: 2015-02-01].

[38] "CoreOS," 2015, URL: https://coreos.com/ [accessed: 2015-02-01].

[39] "Docker," 2015, URL: https://www.docker.com/ [accessed: 2015-02-01].